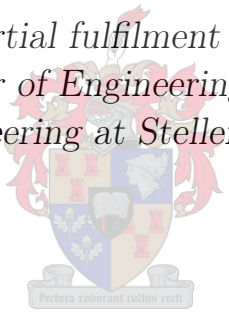
Data Fusion of Radar and Stereo Vision for Detection and Tracking of Moving Objects

by

Frik Botha

Thesis presented in partial fulfilment of the requirements for the degree of Master of Engineering (Electronic) in the Faculty of Engineering at Stellenbosch University



Department of Electrical and Electronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisors:

Dr C. E. van Daalen, Mr J. Treurnicht

March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: ..... March 2017 .....

Abstract

Detection and tracking of moving objects (DATMO) is essential for autonomous navigation systems operating in general environments. Dynamic objects must be identified, localised, and their future positions predicted to assist in decision making regarding path planning and collision avoidance. In addition to its application in autonomous navigation, DATMO also forms the basis of various advanced driver assistance systems (ADASs) that are aimed at making road travel more safe. The research presented in this thesis focuses on the combined use of radar and stereo vision for DATMO. The combination of information from multiple sensors, known as data fusion, introduces redundancy, potentially increasing the confidence and robustness of the system as a whole.

The traditional approach to DATMO is adopted, which involves the chronological steps of measurement extraction, data association and filtering. Measurements are extracted from the radar and vision subsystems independently, using two-dimensional Fourier analysis and sparse feature tracking respectively. A segmentation of moving objects is obtained by a track-to-track fusion algorithm, on data composed of image feature track clusters and Gaussian mixtures originating from radar-based state estimation. Segmented objects are ultimately tracked in a novel implementation of the Gaussian inverse Wishart probability hypothesis density (GIW-PHD) filter that makes explicit provision for extended targets, i.e. targets that generate more than one measurements per time step.

Simulation results indicate significantly improved performance for the proposed data fusion algorithm compared to the case when only vision data is used, due to its increased robustness toward clutter interference. Tests on real-world data do not provide conclusive evidence that suggests improved performance of the proposed radar-vision fusion algorithm compared to vision-only processing. However, practical limitations meant that truly representative datasets could not be gathered. The practical results, however, do indicate very accurate centre point tracking using the GIW-PHD filter, which attests the effectiveness of the Gaussian inverse Wishart model. Moreover, target extent estimates that result from Gaussian inverse Wishart modelling proves sufficiently accurate for the representation of object extent. GIW-PHD filtering also brings about a consistent increase in performance compared to the raw measurements, thereby reinforcing the value of state estimation.

Uittreksel

Deteksie en volging van bewegende voorwerpe is noodsaaklik vir outonome navigasie stelsels wat in algemene omgewings funksioneer. Bewegende voorwerpe moet ondermeer geïdentifiseer en gelokaliseer word. Terselfdetyd moet voorspellings gemaak word oor sulke voorwerpe se toekomstige beweging om besluitneming in verband met padbeplanning en botsingvermeiding by te staan. Benewens die toepassing met betrekking tot outonome navigasie vorm deteksie en volging van bewegende voorwerpe die basis van verskeie gevorderde bestuurdersbystand stelsels. Die navorsing wat in hierdie verslag voorgelewer word fokus op die gesamentlike gebruik van radar en stereo visie vir deteksie en volging van bewegende voorwerpe. Die kombinasie van inligting vanaf verskeie sensore, bekend as data fusie, bring oorbodige inligting teweeg, met die die potensiaal om die algehele gehalte van die stelsel te verbeter.

Die tradisionele benadering tot deteksie en volging van bewegende voorwerpe word toegepas, wat die kronologiese stappe van meting ontrekking, data assosiasie en afskatting behels. Metings word onafhanklik onttrek vir beide die radar en visie substelsels, deur gebruik te maak van twee-dimensionele Fourier analise en yl kenmerk volging respektiewelik. 'n Segmentering van bewegende voorpwerpe word verkry deur middel van 'n toestandsfusie algoritme, wat toegepas word op data bestaande uit groepe beeld kenmerke en radar toestande wat volg uit afskatting. Segmenteerde voorwerpe word uiteindelik gevolg in 'n nuwe implementering van die Gaussies inverse Wishart waarskynlikheidshipotese digtheid afskatter, wat eksplisiet voorsiening maak vir uitgebreide teikens, dit is, teikens wat aanleiding gee tot meer as een meting per tydstap.

Simulasie resultate dui op 'n beduidende verbetering in die prestasie van die voorgestelde data fusie algoritme in vergelyking met die geval waar slegs stereo visie inligting gebruik word, aangesien die metode beter vaar in die teenwoordigheid van steurnisbronne. Toetse op praktiese data gee nie beslissende bewyse om aan te dui dat data fusie beter vaar as die geval waar slegs stereo visie inligting gebruik word nie. Omvattende datastelle kon egter nie ingesamel word nie weens praktiese beperkings. Die praktiese resultate dui egter steeds op baie akkurate middelpunt volging, wat volg uit die toepassing van die Gaussies inverse Wishart waarskynlikheidshipotese digtheid afskatter. Verder lewer die afskatter ook grootte afskattings wat voldoende akkuraatheid bied vir die voorstel van teiken grootte. Gaussies inverse Wishart waarskynlikheidshipotese digtheid afskatting lei ook tot 'n bestendige verbetering in die stelsel uittree in vegelyking met rou metings, wat getuig tot die waarde van afskatting.

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations ...

- Mr Johann Treurnicht, for his guidance and motivation during the course of the project.
- Dr Cornê van Daalen, for his guidance, his help with difficult concepts, and his extremely thorough reviews.
- Clint Lombard, for his part in the development of the data collection platform, his help with dataset collection, and his valued recommendations throughout the project.
- Alex Chiu, for his moral support during the final days, and his lectures on all things probability.
- Aidan Landsberg and Pieter Malan, for the friendly atmosphere in the office.
- Armscor, for their financial support.

Contents

List of Figures

List of Tables

Nomenclature

ACC	adaptive cruise control
ADAS	advanced driver assistance system
CFAR	constant false alarm rate
CPI	coherent processing interval
CPU	central processing unit
CRDM	complex range-Doppler map
CRF	camera reference frame
CUT	cell under test
DATMO	detection and tracking of moving objects
DBSCAN	density-based spatial clustering of applications with noise
EM	electromagnetic
FAST	features from accelerated segment test
FFT	fast Fourier transform
FMCW	frequency modulated continuous wave
GIW	Gaussian inverse Wishart
GIW-PHD	Gaussian inverse Wishart probability hypothesis density
GM-PHD	Gaussian mixture probability hypothesis density
GNN	global nearest neighbour
IF	intermediate frequency
IMU	inertial measurement unit
IQ	in-phase quadrature
JPDA	joint probabilistic data association
KL-diff	Kullback-Leibler difference
KL-div	Kullback-Leibler divergence
LFM	linear frequency modulation
lidar	light detection and ranging
MHT	multiple hypothesis tracking
MOTA	multiple object tracking accuracy
MOTP	multiple object tracking precision
MTT	multi-target tracking
NN	nearest neighbour
OBC	on-board computer

OSPA	optimal subpattern assignment
P3-AT	Pioneer 3-AT
PDA	probabilistic data association
pdf	probability density function
PHD	probability hypothesis density
radar	radio detection and ranging
RCS	radar cross section
RF	radio frequency
RFS	random finite set
RMS	root-mean-square
RRF	radar reference frame
SLAM	simultaneous localization and mapping
SOCA-CFAR	smallest-of cell-averaging constant false alarm rate
sonar	sound navigation and ranging
SPD	symmetric positive definite
SSD	solid-state drive
UKF	unscented Kalman filter
WRF	world reference frame

Mathematical Notation

$\mathbf{X}$	Matrix
$\mathcal{X}$	Random finite set
X	Set
$\mathbf{x}$	Vector
$\mathcal{B}(\cdot)$	Bhattacharyya distance
$\mathcal{N}(\cdot)$	Gaussian distribution
$\mathcal{IW}(\cdot)$	Inverse Wishart distribution
$\mathcal{W}(\cdot)$	Wishart distribution
$p(\cdot)$	Belief distribution
$f(\cdot)$	Dynamic model
$h(\cdot)$	Measurement model
$\hat{(\cdot)}$	Expected value operator
$\otimes$	Kronecker product
$\bar{(\cdot)}$	Mean operator
$\tilde{(\cdot)}$	Scattering matrix operator
$\langle\cdot\rangle$	Set partition operator

Chapter 1

Introduction

Environment perception entails the establishment of spatial and temporal relationships relating a robot and its surroundings. Environmental perception is an important requirement for the autonomous operation of mobile robots and also forms the basis of various advanced driver assistance systems (ADASs). Perception is used for the automation of vehicle navigation, or where potentially dangerous work may be assigned to robots. To navigate reliably, autonomous vehicles should form an understanding of their surrounding scene. In the same manner, driver assistance systems perceive the environment to provide additional safety features to the occupants. Information that result from the perception process serve as input for path planning or collision avoidance algorithms, providing the necessary means for autonomous navigation decision making.

1.1 Overview of Environmental Perception

The perception process is usually divided into two categories, namely: *simultaneous localization and mapping* (SLAM) and *detection and tracking of moving objects* (DATMO) [1]. In SLAM, the purpose is to localise the robot with respect to static objects (landmarks) in the environment. The majority of SLAM applications assume a static environment, or scenarios in which static and dynamic objects can be distinguished. SLAM results in a global pose (position and orientation) estimate of the robot as well as a map containing the position of some landmarks. The map is typically sparse, since most SLAM algorithms focus only on prominent landmarks. SLAM combines motion information with landmark measurements. Motion information is acquired from ego-motion estimation, which pertains to the process of estimating one's own movement. Ego-motion-based navigation in the absence of SLAM is called dead-reckoning. Pose estimates resulting from dead-reckoning are based solely on motion information, and suffers from severe drift due to error accumulation over time. Dead-reckoning navigation naturally extends to SLAM in many permitting environments where adequate landmarks are available. The primary difference between dead-reckoning and SLAM is that in SLAM, landmark locations are maintained in a global map and used to localise the robot more accurately, whereas the former only integrates motion estimates.

DATMO is the detection and tracking of moving objects. Autonomous vehicles may encounter various moving objects with whom they need to interact in a safe manner. For this to be achieved, position and velocity estimates of every moving object in the robot's locale have to be calculated. Some ADAS applications, such as adaptive cruise control

(ACC), require similar information, with the goal of improving the safety of road navigation. DATMO is a fundamental requirement for the task of collision avoidance in both autonomous navigation and ADAS, as it allows for the prediction of object trajectories over time.

DATMO and SLAM are mutually beneficial [2,3]. Knowledge of moving objects enable their exclusion in SLAM calculations, thereby increasing pose estimation accuracy. Moreover, a precise localisation system is essential for moving object tracking from a moving platform [3]. In conjunction, SLAM and DATMO satisfy both the safety and localisation demands of autonomous navigation [3].

Exteroceptive sensors provide the information required for the task of environmental perception. Radars, cameras and lidars are among the common sensors utilised in this regard. These have different properties regarding accuracy, cost, failure cases and more. Any single sensor, however, is inadequate for robust perception in challenging scenarios, prompting the fusion of information from different types of sensors. This practice, known as *data fusion*, can introduce redundancy, potentially increasing the confidence and robustness of the perception system as a whole. The application of data fusion is also encouraged by distinct sensor types that exhibit complimentary characteristics.

1.2 Problem Statement

The work in this project relates to the DATMO facet of environmental perception. In particular, the focus is on data fusion of radar and stereo vision as a means to implement a DATMO system. Combining the information from different sensing modalities adds redundancy and increases estimation accuracy. The goal is to develop a radar-vision fusion approach tailored specifically for DATMO in dynamic, ground-based environments. A moving platform of which the ego-motion is assumed available serves as the carrier for the respective sensors. The system should be able to identify, localise and track any non-stationary object that is within the sensors' field-of-view. Target estimates that are output by the system should be relevant for collision avoidance and autonomous navigation applications.

1.3 Background

The DATMO problem is typically formulated in a Bayesian state estimation framework, involving the chronological steps of data segmentation or measurement extraction, data association and filtering. Methods that proceed in the described manner are categorised as traditional DATMO [4]. Data segmentation aims at the division of sensor data into meaningful pieces such as points, point clusters or line features. This process encompasses the detection part of DATMO, in which the purpose is the identification of moving objects in the environment and the arrangement of the sensor information in a form suitable for subsequent processing.

Multi-target tracking (MTT) follows the measurement extraction process in traditional DATMO, addressing the data association and filtering steps. In MTT, the objective is to determine the number of dynamic objects and their states based on noisy sensor measurements [5]. The data association process entails the assignment of extracted measurements to object tracks; an ambiguous process due to the unknown origin of measurements. Established MTT algorithms include global nearest neighbour (GNN), joint probabilistic

data association (JPDA) [6] and multiple hypothesis tracking (MHT) [7]. The data association process is the only way in which the above mentioned algorithms differ, and is also considered as the most complex aspect of MTT [5, 8]. Associated measurements are filtered in what is the final stage of traditional DATMO. Filtering is the estimation of the underlying state of as system. The result is estimates of the state of moving targets in the environment, e.g. position, velocity and shape.

Variations to traditional DATMO include model- and grid-based approaches [4]. Model-based DATMO avoids the difficult problem of data association by incorporating an object model that dictates measurement to target correspondences. In the setting of autonomous navigation, such models usually include a description of the geometric shape of an object. Grid-based frameworks do not track targets at object level, but rather represent the information using cells into which the environment is discretised. A cell's state is described by a notion of occupancy and velocity. Occupancy grids may be used to implement stand-alone DATMO systems, or to render information that result from traditional DATMO more suitable for path planning and navigation [9].

Achieving reliable performance in real-world environments is a complex challenge in DATMO. Even though the DATMO problem is solved theoretically [3], it cannot shun from hardware inadequacies. Sensors may fail when presented with adverse conditions due to physical limitations or phenomena inherent to their operation. Furthermore, objects may proceed undetected when the available sensors are not suited to detect them. All exteroceptive sensors commonly used in DATMO have undesirable properties. Radar exhibits excellent all-weather operation and is accurate in range, but lacks in its ability to detect targets with low reflectivity. Camera systems provide rich appearance information, are accurate in angle, but perform poorly in ranging applications and are adversely impacted by poor visibility and lighting conditions. Lidars are accurate in both range and azimuth, but suffer in unfavourable weather conditions.

The described limitations serve as encouragement for data fusion. Fusion practices are aimed at circumventing problems arising from sensor deficiencies by exploiting data redundancy. Generally, multi-sensor data fusion offers significant advantages over the use of a single source [10]. The availability of various observations of a single object provides a statistical advantage. Moreover, increased accuracy can be achieved by considering individual sensor properties during fusion.

An integral part of all multi-sensor data fusion algorithms is estimation [11]. The structuring of the estimation process is generally used to distinguish different architectures. Sensory data may be combined at different levels and in a variety of fusion architectures. Techniques for the fusion of raw sensor data typically involve classical detection and estimation methods [10]. In these cases, the problem is usually cast into a Bayesian state estimation framework adapted for multiple sources. Alternatively, fusion may proceed after local processing at each sensor node, or in hybrid configurations containing elements of both.

1.4 Objectives

This thesis presents a radar and stereo vision data fusion method for DATMO. The project objectives are listed below:

1. The main research objective is the fusion of radar and stereo vision information for DATMO. Data fusion should improve the accuracy and robustness of the system.

2. Another primary objective is the development of a state estimation framework. The implementation should allow the states of numerous moving objects to be inferred from the sensor measurements.

3. A secondary objective of the research to perform measurement extraction on the respective sensors' data to detect moving objects. Detection is required in order to enable the demonstration of data fusion and state estimation. Measurement extraction methods should consider generic object classes.

1.5 Project Approach

In this project, information from a short range frequency modulated continuous wave (FMCW) radar and stereo vision cameras are fused for DATMO. Leveraging the angular resolution capabilities of camera sensors may address one of the main shortcomings of radars, while radar range measurements are more accurate compared to that of vision. The detection process is performed for both subsystems individually: a sparse motion-based method identifies moving clusters in image sequences, while radar detection is based on standard Doppler analysis. Separate state estimators are implemented for the respective subsystems. A track-to-track data fusion architecture is developed in which radar and image feature tracks are subsequently combined. Fused estimates serve as measurements in a multi-target tracking framework with explicit provision for extended targets. Figure 1.1 illustrates the processing pipeline of the proposed algorithm.

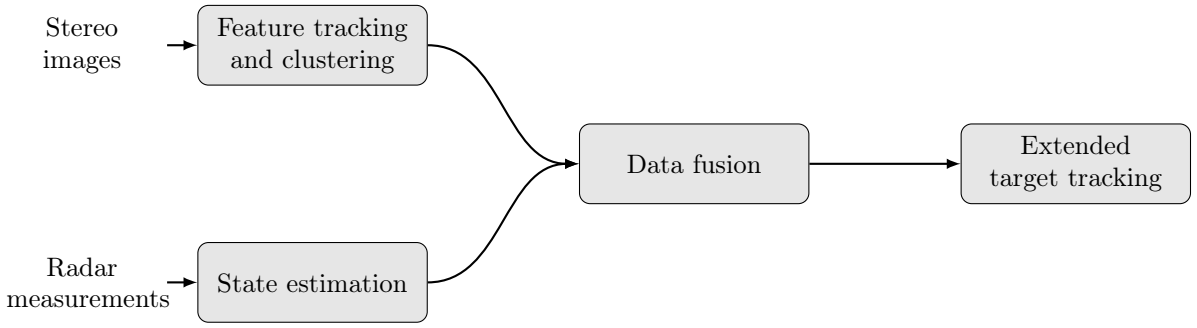


Figure 1.1: Radar-vision data fusion system diagram.

The document is structured as follows. Firstly, a review of relevant research in the fields of DATMO, estimation and data fusion is presented in Chapter 2. Current approaches to combine radar and vision information are also discussed. Chapter 3 describes the physical characteristics that govern measurement generation for radars and stereo vision cameras. The chapter concludes with the presentation of an extrinsic calibration method that is used to determine the relative sensor-to-sensor alignment. Chapters 5 and 6 provide the theoretical foundation for the multi-target tracking framework. The proposed data fusion algorithm is described in Chapter 7. Chapter 8 contains a presentation of relevant results and the analysis thereof. Concluding remarks are given in Chapter 9.

Chapter 2

Literature Review

This chapter details the techniques common to DATMO and data fusion. The goal is to familiarise the reader with the concepts involved in these fields. A discussion of sensors for object detection, including an examination of measurement extraction techniques, is firstly presented. This addresses the detection aspect of DATMO. State estimation and multi-target tracking is subsequently reviewed. Scene and target modelling, as well as data fusion are coupled to the tracking process, and discussed in the following sections. Finally, the current state of the field of radar and vision information fusion is presented.

2.1 Exteroceptive Sensors for Object Detection

In a DATMO setting, exteroceptive sensors gather information that is required for object localisation by surveying the surrounding environment. The following subsections provide a description of the properties and detection methods for different sensor types.

2.1.1 Radar

Radars operate by transmitting radio frequency (RF) electromagnetic (EM) waves and subsequently analysing signals reflected from surrounding objects. By measuring the time delay and phase shift of received signals, the distance and velocity of a target may be determined. The bearing of detections can be retrieved using directional antennas or phase comparison techniques. Sensors following the aforementioned principle of transmission and reception are labelled as active sensors.

In contrast to other exteroceptive sensors, radar's RF waves experience minimal attenuation when penetrating particles such as fog, dust, rain, foliage and smoke [12]. The attenuation caused by these materials is highly dependant on the centre frequency of the waveform, with greater attenuation at higher frequencies [12]. Detrimental to radar-based object detection is its reliance on radar cross section (RCS). The RCS of a target describes its apparent size from the radar's perspective [12]. Large RCS fluctuations may be encountered in a ground-based environment, rendering some objects undetectable.

Radar measurements are generally considered accurate in range, but inaccurate in angle. The angular resolution in ordinary radar operation is determined by the beamwidth of the receive antenna, which in turn is dependant on the centre frequency and physical antenna size. A narrow beam allows more accurate angular measurements, but is not suited to observe large sections of the environment, i.e. to perform volume searches. To

observe a large volume requires either a wide beam, or manual or electronic beam steering. Within-beam angular discrimination is possible when several receive antennas are available, using a technique called monopulse. The monopulse mode of operation enables volume search and angular discrimination requirements to be fulfilled simultaneously.

Measurement error may arise due to multipath effects. These errors occur when the wave propagates back to the radar by more than one path. The increase in propagation distance delays the signal, creating the impression of a false target that is at a greater distance [13]. Multipath increases with wavelength, making radar more likely to suffer in comparison with sensors operating in or near the light spectrum [12].

Millimetre wave radar is the preferred candidate in the field of robotics since applications are often subjected to size and mass constraints. This class of radars operate in the range of frequencies between 30 GHz and 300 GHz. The short wavelength of millimetre wave radars is its primary advantage, allowing the use of physically small antennas [12].

2.1.2 Sonar

The operating principle of sonar is similar to that of radar. Sonar relies on mechanical instead of EM waves, thus requiring a physical medium for signal propagation. The physics of acoustic sensing is not favourable for environment perception [14], and it therefore sees limited use in DATMO applications. All solid surfaces are acoustic reflectors, and most surfaces display mirror-like (specular) acoustic behaviour [14]. The consequence of specular scattering is that angled surfaces reflect the incoming signal away from the source, thereby hindering detection [14]. Specular scattering may also introduce multipath ranging errors [9]. Sonar is best suited to underwater applications, due to reduced attenuation [15] and since other exteroceptive sensors are incapable of operating in water.

2.1.3 Lidar

Lidars are among the most commonly used and versatile sensors for perception applications. Like radar and sonar, lidar measurements are in range-bearing form. Lidars retrieve distance by measuring the time-of-flight of a transmitted light beam [4]. The sensor exhibits excellent angular resolution and accurate range measurements, enabling lidar's use in high performance perception systems. Two-dimensional lidars return very sparse measurements much like narrow beam radars. Three-dimensional lidars are able to perform volume searches, but they are very expensive.

Backscatter from weather phenomena such as rain, fog, dust or snow may result in unwanted detections [16]. Lighting conditions may also hinder reliable detection, e.g. light absorbing material or the presence of direct sunlight [9]. The vulnerability of lidar to external factors is its most significant deficiency.

2.1.4 Vision

The abundance of cheap camera sensors has contributed to the extensive use of computer vision for environmental perception. Camera systems are versatile, unobtrusive, and the development cycle is easily initiated with inexpensive commercial components. Cameras provide very accurate angular measurements and, for stereo configurations, coarse range. Semantic information may also be extracted in addition to distance and angle. The influence of external environmental factors on vision systems are generally the same as for lidars.

The wealth of information provided by cameras allow for a multitude of different approaches to obstacle detection, permitting a more detailed discussion. The following sections detail the most prominent vision-based moving object detection methods. For the purpose of this discussion, the terms *object detection* and *segmentation* will be used interchangeably.

Background Subtraction

Background subtraction is perhaps one of the simplest and most intuitive methods of motion segmentation in images. It amounts to pixel-wise subtraction of a background model from the current frame, thereby emphasizing areas where the image has changed. Background subtraction assumes a stationary camera and is therefore not suitable for use on a moving platform.

Attempts have been made to perform background subtraction for moving cameras [17, 18]. In these methods motion induced by camera movement has to be corrected for. This relies on accurate estimations of ego-motion which is attainable through the use of an inertial measurement unit (IMU). Ego-motion may also be estimated by using optical or range sensors.

Geometric Measurements

Certain assumptions allow general physical attributes to be used for the detection of generic object classes. Examples include texture, edges, shadows and symmetry information. It is reasonable to expect that arbitrary objects will contain some combination of geometric features, although some are more restrictive than others; for example symmetry. A downside of this approach is that stationary clutter and targets of interest are indistinguishable. Researchers therefore usually limit the implementation to areas where motion is assumed [19, p. 346]. In an ADAS environment for instance, the road surface would first be detected, and the result used to only consider geometric shapes that are above the road surface.

Optical Flow

An important cue in image analysis is the relative movement of visible surfaces between successive frames. Two-dimensional image plane motion fields, or optical flow, provide important information regarding the spatial arrangement of a scene. Various techniques exist for computing optical flow, including differential-, energy-, correlation- and feature-based [20]. Optical flow may be computed for either the entire image (dense) or for selected patches only (sparse). The addition of range data to optical flow calculations result in three-dimensional motion fields called *scene flow*.

Differential optical flow is the most widely used technique, and is based on the assumption that the intensity of a moving pixel remains constant over time, i.e.

$$I(u, v, t) = I(u + \Delta u, v + \Delta v, t + 1) \tag{2.1}$$

where $[\Delta u, \Delta v]^T$ is the flow vector of pixel $[u, v]^T$ from time t to $t + 1$ [21]. The flow vector may be solved by optimizing the Taylor series expansion of Equation (2.1), often by means of the Lucas-Kanade method [22].

Differential optical flow can be categorised into local and global methods according to the type of energy function they optimise. Local optical flow methods are relatively

robust under noise, but do not result in motion estimates at every pixel [21]. Smoothing constraints allow dense motion fields, and are usually formulated in terms of a global optimisation problem containing local (data) and smoothing terms. Textureless regions is problematic for both local and global optical flow calculations [23]. Local methods generally disregard such regions, while the smoothing constraints in global methods enable their resolution.

Discontinuities in the motion field can assist in the segmentation of an image into regions that correspond to different objects. Various authors have used this principle for generic moving object detection in image sequences [24–26]. Klappstein et al. [24] demonstrates this principle using both monocular and stereo vision. By introducing motion constraints, features in the image are either considered as static or belonging to moving objects. Features are then clustered into coherent objects and subsequently used as seeds for segmentation using graph cuts. A similar approach is followed by Wedel et al. [25] using dense scene flow.

Appearance-Based Methods

Appearance-based methods involve the implementation of machine learning techniques to learn semantic patterns in images. A common approach is supervised learning, which entails the training of a classifier using manually labelled data. A drawback of these methods is that classifiers are class specific. Multiple object detection would require as many classifiers, which is often impractical due to the computational burden. Learning methods do not require ego-motion estimates and are robust against irregular movement of the sensing platform.

The following paragraph motivates the choice of radar and vision for data fusion. As was stated earlier, the performance of sonar is not up to standard in air. Lidars take accurate measurements, but an expensive 3-D lidar is required for volume searches. Vision, and more recently also radar, are the affordable options for data fusion applications. Both are able to do volume searches, and their complimentary characteristics favour data fusion. Moreover, radar-vision fusion may be extended to small airborne platforms since both sensors are lightweight.

2.2 Recursive Bayesian State Estimation

The eventual purpose in DATMO is to extract the states of moving object using the information from exteroceptive sensors. The standard approach in which these quantities are obtained relies on stochastic modelling of the processes surrounding target dynamics and measurement generation. The Bayes filter provides a rigorous theoretical foundation to infer target states using the above mentioned probabilistic models.

In a Bayesian context, the desired target states are represented as a probability distribution $p(\cdot)$, also know as the *belief*. The task is to estimate, at each time step, the current states of a target given all sensor measurements that have been collected, i.e. the *posterior distribution* $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, where $\mathbf{x}_k$ is the state vector at time k, and $\mathbf{z}_{1:k}$ is all target measurements up to and including time k.

The Markov assumption allows the elegant recursive formulation of state estimation, i.e. the Bayes filter. In a Markov process, the dynamic model is assumed to be independent

of all previous states given the present state, i.e.

$$f(\mathbf{x}_k|\mathbf{x}_{1:k-1}) = f(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{2.2}$$

The Bayes filter consists of two steps, namely prediction and correction. Prediction describes the transition of the belief distribution from the previous time step to the current without additional observations of the target. Evaluating the prediction step results in the *prior distribution*, and requires a model of the target's dynamic behaviour. Formally, the prior is given by [9, p. 27]

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int f(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1}, \tag{2.3}$$

where $f(\mathbf{x}_k|\mathbf{x}_{k-1})$ is a probability distribution modelling the target dynamics. Sensor information is incorporated in the correction, or measurement update, step. The required posterior can be calculated according to Bayes' rule as

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{h(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{\int h(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})d\mathbf{x}_k}, \tag{2.4}$$

where $h(\mathbf{z}_k|\mathbf{x}_k)$ is the sensor or measurement model, which is a conditional distribution modelling the measurement generation. The denominator of Equation (2.4) is simply a constant and can be considered a normalisation factor.

Equations (2.3) and (2.4) define the single-source, single-target Bayes filter, which forms the theoretical foundation for single-target tracking, single-target information fusion, as well as multi-sensor and multi-target detection, tracking and data fusion [27].

2.3 Multi-Target Tracking

En route to implementing a practical target state estimator, one of the fundamental assumptions of the Bayes filter needs to be addressed, namely perfect measurement-to-track associations (hence the term 'single-source, single-target'). The environments under consideration for DATMO fail to comply with this assumption. A typical MTT scenario is shown in Figure 2.1. The expected observations at time k, $h(\hat{\mathbf{x}}^{(i)}_{k|k-1})$, are drawn along with their validation gates. The state vector $\mathbf{x}^{(i)}_{k|k-1}$ describes the track of the i^{th} target. A gate defines the region where measurements will be considered for association to the particular target track, and is a concept utilised in many multi-target tracking (MTT) algorithms. The presence of multiple targets and clutter entail uncertain relationships between measurements and sources: an individual measurement may have originated from any one of the targets or from static clutter in the environment. Each incoming sensor report needs to be assigned to the correct target track to facilitate robust estimation. Ambiguous data association is the primary motivation for MTT techniques. The remainder of this section will present some established and more recent methods that address data association, which is the defining component of different MTT algorithms.

2.3.1 Global Nearest Neighbour

The simplest and most intuitive data association method is the nearest neighbour (NN) algorithm. NN assigns the measurement closest in statistical distance to the predicted track to be used in the measurement update. The NN algorithm is in fact a single-target

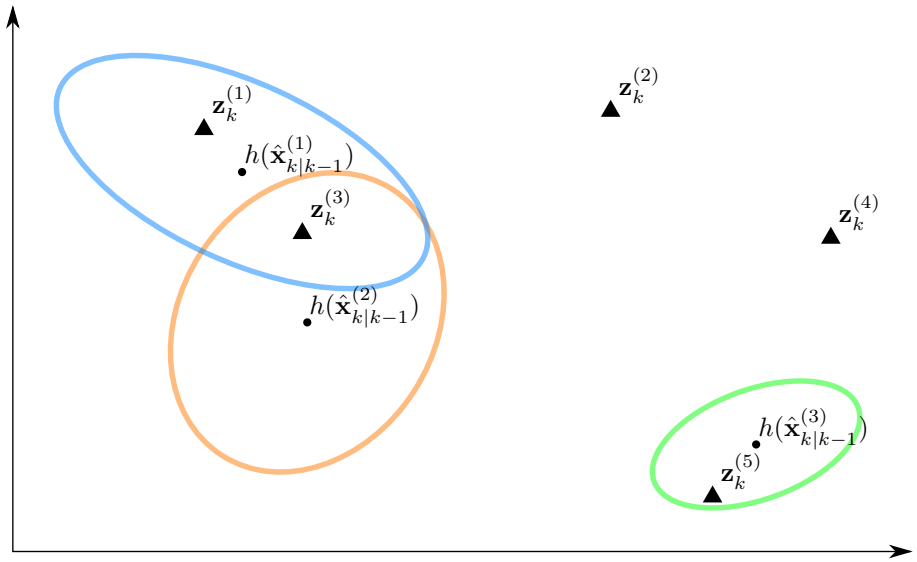


Figure 2.1: Target validation gates in a typical multi-target tracking scenario. A gate is centred on the expected observation position $h(\hat{\mathbf{x}}_{k|k-1}^{(i)})$, and is proportional to the innovation distribution. Measurements that fall within a target's gate are considered for subsequent data association.

data association method, since surrounding targets and their associations are not taken into regard when searching for the nearest measurement. Its extension to multiple targets is knows as global nearest neighbour (GNN).

GNN makes a hard assignment of at most one observation to a track when measurements arrive at any given scan time [5]. In contrast to single-target nearest neighbour data association which is only locally optimal, in GNN, distances between all measurements and tracks are considered when calculating the association assignments. The eventual assignment mapping is chosen as the most likely solution among all mappings [28]. Clutter measurements or new targets account for unassociated measurements, which proceed to initiation logic in order to form new target tracks.

2.3.2 Joint Probabilistic Data Association

Fortmann and Bar-Shalom [29] introduced the joint probabilistic data association (JPDA) filter, which is the multi-target extension of single-target probabilistic data association (PDA) filter. Both standard PDA and JPDA are non-optimal in the sense that they require Gaussian approximations of the posterior distributions [30]. In PDA, all measurements that are within the target's validation gate are considered in the measurement update step. This is implemented by substituting the traditional single-measurement innovation with a weighted combination calculated from all gated measurements. The weighting factors are calculated in accordance with the likelihood of the particular measurement. Such probabilistic measurement-to-track associations of numerous observations are known as soft assignments. Soft association techniques are slightly more complex than their hard assigning counterparts, but they prove much more effective in cluttered environments.

The key difference in multi-target JPDA, as opposed to single-target PDA, is the way in which the mixture weights are calculated [28]. The individual measurement likelihoods used in determining the weights are computed with all the targets and measurements in

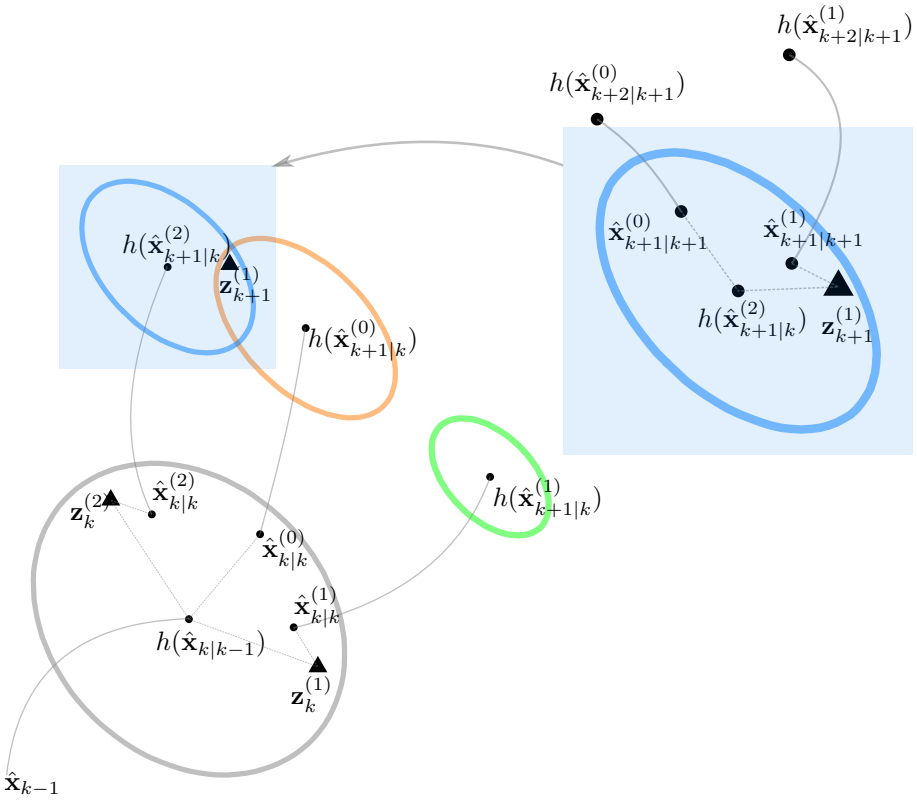


Figure 2.2: Progression of the multiple hypothesis tracking algorithm. At each time step, a hypothesis is created for all measurements that fall within the validation region of a target. An additional hypothesis to account for missed detections is also initialised. Grey lines indicate either a prediction or measurement update step. The enlarged blue ellipse is used to show the second update recursion for the particular target. Figure adapted from Durrant-Whyte [31].

consideration. The process can be quite complex and is also dependant on an assumed clutter distribution. Upon calculating the weights, the JPDA algorithm proceeds in the same manner as the PDA filter [31].

2.3.3 Multiple Hypothesis Tracking

The preceding data association approaches consider target states of a single time step in their calculations. All previous information is encapsulated in the probability distributions of the respective targets. The multiple hypothesis tracking (MHT) algorithm introduced by Reid [7] specifies a framework that regards target states across multiple scan times. The MHT filter initialises separate tracks, or hypotheses, for every measurement in a target's validation gate. In addition, a hypothesis is also generated to represent the missed detection/false alarm case. The indiscriminate associations result in a 'tree' of hypotheses being spawned for a target at every time step. The likelihood of a particular branch of the tree is given by the recursively evaluation of the association likelihoods along the branch. To limit the combinatorial growth in track hypotheses, a pruning strategy based on the likelihoods is usually implemented [31]. The partial tree structure and hypothesis generation of a single target over two time steps is shown in Figure 2.2.

The MHT algorithm described above and illustrated in Figure 2.2 is known as track-oriented MHT, which is one of the most advanced MTT methods [5]. MHT is generally the

apt choice in situations with high track uncertainty, e.g. crossing or manoeuvring targets, and is also effective in clutter [5, 31]. Practical implementation of the method requires careful consideration with regards to pruning in order to limit memory consumption.

2.3.4 Probability Hypothesis Density Filter

Mahler's [32] random finite set (RFS) modelling permits the proper Bayesian formulation of multi-target recursive filtering. In this scheme, both target states and observations are modelled as random sets consisting of a random number of random elements. RFS modelling stems from the intuitive finite set form of representation for a collection of targets and measurements, i.e.

$$X_k = \left\{ \mathbf{x}_k^{(i)} \right\}_{i=1}^{N_{\text{targets},k}}, \tag{2.5}$$

$$Z_k = \left\{ \mathbf{z}_k^{(i)} \right\}_{i=1}^{N_{\text{measurements},k}}, \tag{2.6}$$

where the elements of X_k represent individual target state vectors, and those of Z_k individual observation vectors. In this setting, uncertainty in target states is characterised by modelling the state and observation finite sets as random finite sets, in analogy to the random vector modelling in single-target systems.

The theoretical development lead to the multi-target equivalent of the single-target Bayes filter. As is the case for the single-target variant, the resulting recursive equations are intractable. Approximations involving the propagation of lower order statistical moments have been implemented by Mahler [32] and Vo [33]. These lower order moments are termed *probability hypothesis densities*, and the resulting recursive filter the *probability hypothesis density* (PHD) filter.

The key advantage of RFS-based recursive filtering is the avoidance of the data association problem. The algorithm does not implement any form of measurement-to-track assignments, due to the set-based target and observation modelling. In the PHD filter, explicit track formation is avoided and exchanged for a representation that describes the probability of a target being present in a certain space [5]. Consequently, target identity tracking is not incorporated in the standard PHD algorithm.

2.4 Measurement Modelling

The discussion in Section 2.2 detailed the underlying theory upon which target tracking is built. As was mentioned, the Bayes filter recursion requires dynamic and measurement models for the posterior states to be inferred. This section will discuss common modelling methods with regards to the measurement model. These models describe the distribution of sensor measurements conditioned on the target state.

The standard sensor model used in multi-target tracking algorithms assume negligible object extent by modelling a target as a point in space [34]. The premise of a DATMO or ADAS application entails that objects appear 'near' the sensing platform. As a consequence, an object will most often occupy multiple sensor resolution cells and generate numerous measurements per time step, thereby violating the point target assumption. Tracking these objects give rise to the problem of *extended target tracking*. Formally, an extended object is defined as one which may generate a varying number of spatially distributed measurements per scan [34]. An illustration of an extended target is given in

Figure 2.3. Object shape, as well as the expected number of measurements per object can be incorporated in extended target models.

Related but distinct to extended target tracking is the problem of group tracking. Group targets consist of a number of a objects that move in a coordinated and interacting fashion [35]. Extended target tracking techniques are often directly applicable to group tracking due to the similarities that distinguish these problems from conventional point object tracking. A notable difference is the interacting nature of groups that is not present in extended targets. The discussion will proceed with emphasis on extended targets because of their prevalence in DATMO and ADAS applications.

Extended target models usually assume a specific geometric shape. The parameters that describe such a shape are appended to the target state vector and inferred from observations. For this to be achieved in the Bayesian filtering paradigm requires the specification of an augmented sensor model that explicitly accounts for the target extent. Common geometric models include curves in 2D or 3D space, ellipses or rectangles.

2.4.1 Spatial Distribution Models

Gilholm and Salmond [36] developed the proper Bayesian formulation for extended target tracking, in which the target is represented by a spatial probability distribution. In addition to the spatial distribution, the number of measurements originating from a target is modelled as a Poisson process [36]. The spatial model represents the distribution of measurement sources across the target, and may take the form of general shape models such as lines or circles. Measurements are assumed to be independent realisations from the spatial probability model convolved with a sensor error model [36]. Therefore, the probability density function (pdf) of a single target measurement, as described by the spatial model, is given by

$$h(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}|\mathbf{y})p(\mathbf{y}|\mathbf{x})d\mathbf{y}, \tag{2.7}$$

where $p(\mathbf{y}|\mathbf{x})$ denotes distribution of measurement sources conditioned on the target state $\mathbf{x}$, and $p(\mathbf{z}|\mathbf{y})$ denotes the distribution of an individual measurement $\mathbf{z}$ conditioned on the measurement generating sources $\mathbf{y}$. The eventual likelihood is a combination of Equation (2.7) and the Poisson distribution that models the expected number of measurements. The inclusion of a geometric model leads to a significantly more complicated sensor model

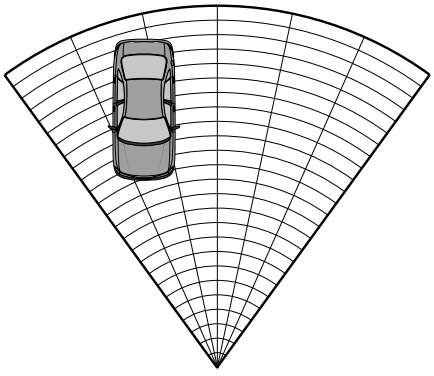


Figure 2.3: Illustration of an extended target. An extended target occupies multiple sensor resolution cells and may as a result generate numerous spatially distributed measurements per scan.

compared to the point target case. As a result, the implementation is usually restricted to particle filters [36].

2.4.2 Random Hypersurfaces

An extended target algorithm based on the concept of a random hypersurface was introduced by Baum and Hanebeck [37]. Any individual measurement is modelled as a noisy observation of a measurement source on the target surface [37]. A measurement source is assumed to be a randomly generated hypersurface representing a scaled version of the target shape. A two step process therefore defines the generation of measurements: first, a measurement source model is generated, after which the sensor error model describes the observations to be expected from the particular source. The likelihood of an individual measurement is determined by evaluating the combined probability of the measurement source given the current shape parameters, and the measurement given the source. Various shapes can be used with random hypersurface models, including ellipses and arbitrary star-convex[1] patterns [37, 38].

The random hypersurface algorithm does not explicitly model the position of measurement generating sources, but rather estimates the shape of the target [38]. Explicit source models are common in extended target tracking. However, they require proper models of the measurement generating sources and sensor resolution that is high enough for the different sources to be resolved [38]. In addition, explicit source models may suffer when sensor returns are influenced by the target's orientation to the sensor.

2.4.3 Random Matrices

Koch [39] formulated a very efficient model for extended target tracking based on the use of symmetric positive definite (SPD) random matrices. Each measurement is interpreted as a measurement of the object's centroid with an error that is proportional to the object's extent [39]. By this procedure, the extent is incorporated into the measurement likelihood function as a covariance substitution, allowing very simple update equations. SPD matrix representation implies an elliptical object shape, and is modelled using an inverse Wishart distribution. The inverse Wishart distribution is the conjugate prior for the unknown covariance matrix representing the measurement spread [34,39]. The novelty of the method, in contrast to previous elliptical shape modelling, is the joint estimation of centroid kinematics and physical extent in a rigorous Bayesian framework [40]. A shortcoming of Koch's original method is the neglect of any statistical sensor error. Feldmann et al. developed an adapted random matrix algorithm that takes both the object extent and sensor error into account when modelling the spread of measurements [41]. The formulation is similar to Koch's, and the shape representation is exactly the same. Without the inclusion of the sensor error, the algorithm would effectively estimate object extent plus sensor error.

In both random matrix approaches, the extent is updated by a weighted combination of matrices representing the predicted extent, measurement scattering and the mean measurement deviation respectively [39, 41]. The measurement update formulas derived for the random matrix algorithm are considerably simpler than those resulting from spatial distribution or random hypersurface modelling, and may be implemented in a linear filter. A limitation of the random matrix model is the elliptical shape constraint.

[1] A set $S \subset \mathbb{R}^N$ is star-convex with center $\mathbf{m}$ if each line segment from $\mathbf{m}$ to any point in S is contained in S [38].

2.4.4 Alternative Modelling Approaches

The three extended target tracking algorithms described above make up the current state of the field in Bayesian extended target tracking. Some methods also proceed in a non-Bayesian fashion due to the difficulty of formulating tractable measurement models for extended targets. This is especially common in computer vision applications, where 'extended' objects are often tracked using appearance information to match shape kernels between frames [42]. Such approaches are fundamentally different to Bayesian extended target methods, in which shape is inferred over multiple time steps.

The preceding extended target tracking techniques all rely on a state vector representation for individual targets, in accordance with traditional or model-based DATMO. Grid-based alternatives may also be used to represent the environment. Rather than storing moving object information in separate state vectors, grid methods rely on a division of the environment into discrete cells. Typical properties that describe each cell, such as velocity and/or occupancy, are subsequently estimated from sensor measurements. In the context of advanced driver assistance system (ADAS), a local grid is typically used to represent the area in front of the vehicle [4, 43]. The grid then serves the purpose of modelling potential dangers rather than building and maintaining a global map [4].

2.5 Data Fusion

The previous sections introduced some of the important concepts relating to Bayesian state estimation and target tracking. The discussion now moves on to data fusion, which has the potential to improve the confidence and robustness of tracking systems.

Multi-sensor data fusion seeks to achieve superior inferences by combining information from multiple sources [30]. Safety critical systems need to be robust with respect to adverse environmental conditions, and can therefore not dependant on single-source information. Complimenting characteristics of different sensor types may improve the overall observability of physical target attributes. Multiple reports from comparable sources also improves estimation performance, and is analogous to multiple observations from a single sensor [30, p. 2]. This section introduces the core methods and concepts applicable to data fusion in a DATMO application.

2.5.1 Probabilistic Data Fusion

Optimal fusion in the Bayesian sense relies on the processing of multi-sensor information at a centralised state estimator [5]. Data fusion is then formulated in the classical Bayesian methodology adapted for multiple sources. Consider the synchronous set of observations from $N_{\mathrm{s},k}$ sensors at scan time k

$$Z_k = \left\{ \mathbf{z}_k^{(i)} \right\}_{i=1}^{N_{\mathrm{s},k}}. \tag{2.8}$$

It is desired to use all the available information to construct the posterior $p(\mathbf{x}_k|Z_k)$. Direct implementation of Bayes rule results in

$$p(\mathbf{x}_k|Z_{1:k}) = \frac{h(Z_k|\mathbf{x}_k)p(\mathbf{x}_k|Z_{1:k-1})}{\int h(Z_k|\mathbf{x})p(\mathbf{x}|Z_{1:k-1})d\mathbf{x}}, \tag{2.9}$$

which is the multi-sensor equavalent of Equation (2.4). Due to the difficulty in constructing the joint sensor model $h(Z_k|\mathbf{x}_k)$, it is usually assumed that respective sensors generate

readings independently from one another given the state $\mathbf{x}$, i.e. [31]

$$h(Z_k|\mathbf{x}_k) = h\left(\mathbf{z}_k^{(1)}, \ldots, \mathbf{z}_k^{(N_{\mathrm{s},k})}|\mathbf{x}_k\right) = \prod_{i=1}^{N_{\mathrm{s},k}} h(\mathbf{z}_k^{(i)}|\mathbf{x}_k). \tag{2.10}$$

Equations (2.9) and (2.10) provide the theoretical grounds for optimal Bayesian multi-sensor data fusion. This formulation of data fusion entails the derivation of appropriate sensor models and application of the above equations in a recursive state estimator. Sensor reports are, however, seldom synchronous in real-world scenarios. Fusion may then be conducted by means of separate updates for the respective sensors [44]. Bayesian multi-sensor data fusion is directly applicable to traditional state space models as well as probabilistic grids [11].

2.5.2 Feature-Level Fusion

Bayesian data fusion is not possible in the event that sensors do not measure the same physical phenomena [30]. Sensors are therefore required to report comparable information if the multi-sensor Bayes filter is to be applied. Feature-level fusion entails the extraction of representative features from the sensor data in order to obtain compatible information from numerous sources that may be dissimilar. Pattern recognition techniques such as regression or clustering algorithms can subsequently be implemented on a multi-source feature vector for improved classification or decision making [30].

Both the feature-level and Bayesian techniques are categorised as *centralised fusion*, since these methods require low-level sensor data at the fusion centre. Centralised fusion architectures demand high communication bandwidth between individual subsystems and the central processing unit, but is generally considered superior to alternative fusion methodologies.

2.5.3 Track-to-Track Fusion

The previous data fusion techniques require low-level sensor information. Low communication bandwidth or restrictions of legacy sensors may render such data unavailable. Generally, the only available information in these scenarios is the distribution parameters describing target tracks, and combining the information amounts to track-to-track fusion. In classical track fusion, each sensor generates tracks independently from other sensing nodes, before transmitting the pre-processed information to a central processor [30]. Here, tracks are associated and their information combined to acquire more accurate target information.

Data fusion strategies cannot be inhibited to a select number of algorithms, since they can be devised in countless ways. Hybrid configurations between centralised and track-to-track fusion architectures may be implemented, where a combination of pre-processed and raw data is used [30]. Other architectures include distributed fusion, in which sensor nodes are interconnected without the concept of a central processor. However, for the purpose of exteroceptive sensor fusion for DATMO applications, lower level fusion architectures are most applicable.

2.6 Radar-Vision Data Fusion

The foregoing sections laid the groundwork for the analysis of current approaches in radar-vision data fusion. Radar and camera sensors exhibit complimentary sensing characteristics that may prove beneficial to combine for DATMO and other environmental sensing purposes. Leveraging the angular resolution capabilities of camera sensors may address one of the main shortcomings of radars, while radar range measurements are more accurate compared to that of vision. The error bounds of these respective sensors are shown in Figure 2.4. The aim is to reduce the estimation error to fit the best of both sensors by proper combination of their information. What follows is an exposition of the state of the field with regards to radar and vision data fusion for DATMO.

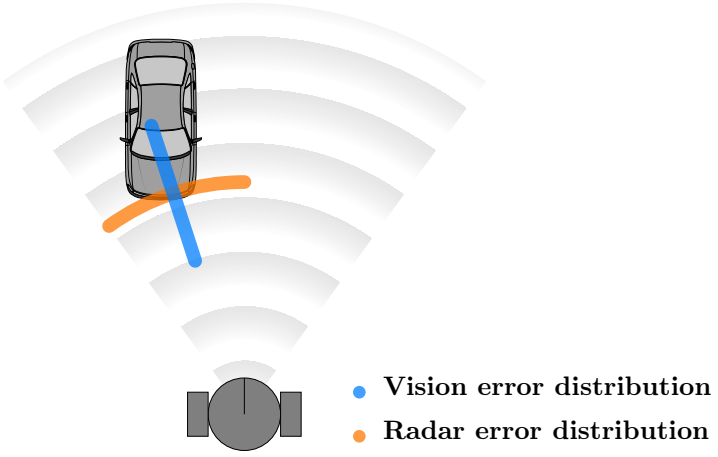


Figure 2.4: Vision and radar error characteristics.

2.6.1 Attention Windows

The concept of attention windows is encountered quite often in literature regarding radar-vision fusion [45–48]. In these methods, the radar provides a list of targets which are subsequently processed using image data to refine, validate or classify the target. Radar detections therefore serve as a guide to subsequent image processing algorithms.

Gern et al. [45] demonstrates such an approach to improve the lateral information of an object. Different vision-based operators are defined that vote individually for the centre of a radar attention window. A decision module eventually determines the extent of the object using the output of the operators. Wang et al. [46] follows a similar approach, but instead uses edge and shadow information for determining object extent. An alternative involving machine learning is presented by Ji and Prokhorov [47]. Their approach forwards attention window information to an in-place learning framework to classify objects. The radar effectively reduces the search space of the classification framework, allowing real-time performance.

2.6.2 Fusing Independent Observations

Work where independent observations from the radar and vision subsystems are fused are few in number. Wu et al. [49] demonstrates a feature-level fusion algorithm which

constructs refined object contours using information from a radar and stereo vision cameras. Measurement extraction is initially performed independently for the separate subsystem: the radar tracks point targets, while two-dimensional contours are extracted from the stereo depth maps. Subsequent feature-level fusion involves the association of radar tracks to image contours, and their eventual refinement to obtain improved range and azimuth estimates. Fused contours are ultimately tracked in an extended target tracking framework [49]. Their extended target model does not adhere to proper Bayesian principles, namely joint kinematic-extent estimation, but rather tracks the contour parameters independently.

A solution that borrows from the attention window methodology is presented by Fang et al. [50]. Their algorithm starts by calculating edge maps from stereo image pairs. Edge pixels are subsequently arranged into histogram bins according to their disparity values. Peaks in the resulting histogram serve as hint that an object may be present, and is used to define disparity ranges for subsequent object segmentation. This allows the implementation of a multi-level segmentation procedure based on different disparity intervals. Information describing the number of targets and their distances is optionally incorporated in the specification of depth ranges for vision-based segmentation [50]. Radar detections then essentially guide the multi-level object segmentation.

Richter et al. [51] developed an environmental sensing system that detects both stationary and moving objects, using ego-motion data along with information from a monocular camera and a radar. The detection process is again executed individually for the subsystems. In this case, vision-based detection relies on u-shape template matching. These detections are used to verify radar observations and to estimate additional target properties such as width and lateral position [51].

Relatively few sources address the problem of combining the information of radars and cameras. Related work tends more towards lidar-vision fusion due to the rich and consistent information extracted by lidars. A common trend among radar-vision fusion literature in short-range applications is the rarity of mathematical techniques such as multi-sensor Bayesian filtering. The apparent reason is the fairly high dissimilarity of the information provided by these sensors when applied in ground-based DATMO environments, leading researchers to adopt either feature- or track-based data fusion methods.

A notable shortcoming in all the prominent radar-vision fusion sources referenced above is the lack of probabilistic modelling of object extent. The majority of these research efforts focus on radar-guided image processing [45–48,50], feature-based fusion [49] or cross validation of detections [51,52]. Although geometric information is often estimated [49,51], kinematic and extent estimation does not follow the principles laid out in Section 2.4, but rather regards the two as decoupled entities. In addition, multi-target tracking is generally given little attention [45–47].

The work in this project seeks to address the above mentioned shortcomings. Explicit extent modelling and the incorporation thereof in joint kinematic-extent tracking is expected to improve estimation accuracy. Embedding extended target models in advanced MTT algorithms should also contribute to an increase in perception performance.

Chapter 3

Sensor Configuration

This chapter sets out to describe the hardware involved in the project. Specific sensors have been acquired for the implementation of the proposed DATMO system, and knowledge of their operation will aid the subsequent understanding of some hardware specific solutions implemented in the project. To begin, an overview of the components that constitute the system will be presented. A thorough discussion with regard to the physical operation of the respective sensors will follow. Finally, the calibration method developed for geometric alignment of the radar and vision subsystems will be presented.

3.1 Hardware Overview

This project's sensing hardware consists of two Point Grey Flea3 cameras and a short-range radar. The sensors are connected to the on-board computer (OBC) of a Pioneer 3-AT (P3-AT) mobile robot to which they are rigidly mounted as shown in Figure 3.1. The P3-AT's OBC has a dual core processor and runs 64-bit Ubuntu Linux. For the purpose of this research, the robot platform was used to gather datasets for offline processing. A solid-state drive (SSD) accompanies the processor to facilitate real-time data storage.

The radar utilised in this work is an Infineon BGT24MTR12 development kit unit, with a centre frequency of 24 GHz. The radar has one transmit- and two receive patch antennas as illustrated by the checkerboard-like patterns in Figure 3.1. Dual receive antennas allow angular discrimination in one spatial dimension by the principle of *monopulse* (see

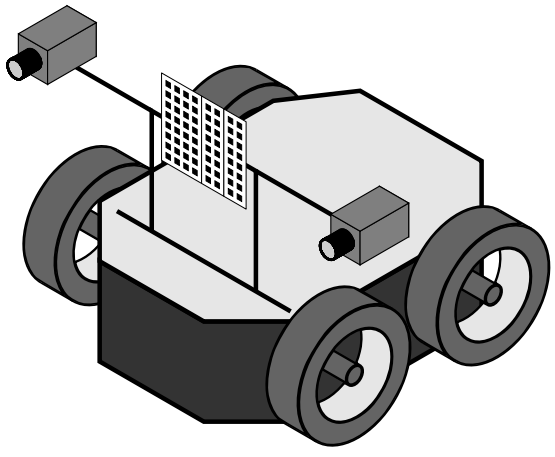


Figure 3.1: Sketch of the sensing platform used in this project.

Section 3.2.3). The radar is operated according to FMCW principles and at a bandwidth of 155 MHz. Analog signals from the respective receive antennas' in-phase quadrature (IQ) receivers are synchronously sampled to render complex-valued information.

The cameras are mounted in a horizontal stereo configuration on either side of the radar as shown in Figure 3.1. The stereo baseline is approximately 55 cm, which is nearly the same as that of the renowned KITTI dataset [53]. A hardware trigger ensures time synchronisation between the respective cameras. Synchronous triggering and global shutters allow precise stereo image pairs to be captured. The cameras are operated in monochrome mode, at a resolution of 1280×960 pixels. Camera recording is set at 10 Hz, while the radar achieves a marginally faster rate.

Both subsystems stream unprocessed data to the OBC. For the radar, this data is in the form of sampled data representing the complex valued signals of the two receive channels. Images are transmitted as 8-bit RAW format. A global CPU timer assigns a time-stamp to every measurement upon its arrival.

Before advancing, a mention of the relevant coordinate systems is in order. With reference to Figure 3.2, three reference frames can be defined. The common frame of reference in which estimation will be performed is the world reference frame (WRF), denoted by W. A separate frame for the robot is not chosen. Instead, the robot's position is assumed to coincide with the centre of the camera reference frame (CRF) C. The radar reference frame (RRF) is denoted by an R. The chosen axes orientations coincides with the usual stereo vision coordinate frame, with the Z axis pointing out of the camera. The relative alignment between the RRF and CRF is the subject of Section 3.3.

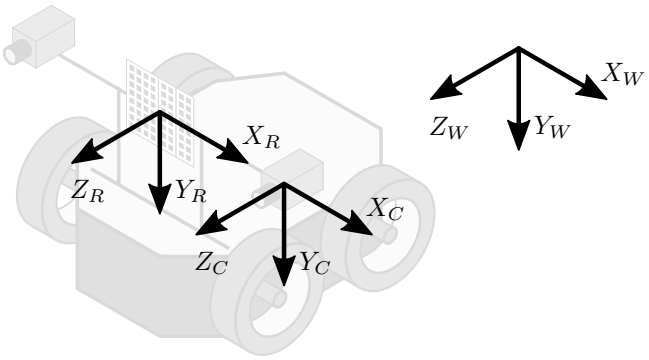


Figure 3.2: Coordinate system conventions.

3.2 Physical Sensor Models

In this section, the physical phenomena that governs the operation of camera and radar sensors will be examined. An understanding thereof is necessary to grasp subsequent discussions on measurement extraction and sensor calibration.

3.2.1 Pinhole Camera Model

Relating camera information to the physical world requires models that describe the mathematical relationship between 3D world coordinates and image coordinates. The most simple of such models in computer vision is the pinhole camera model.

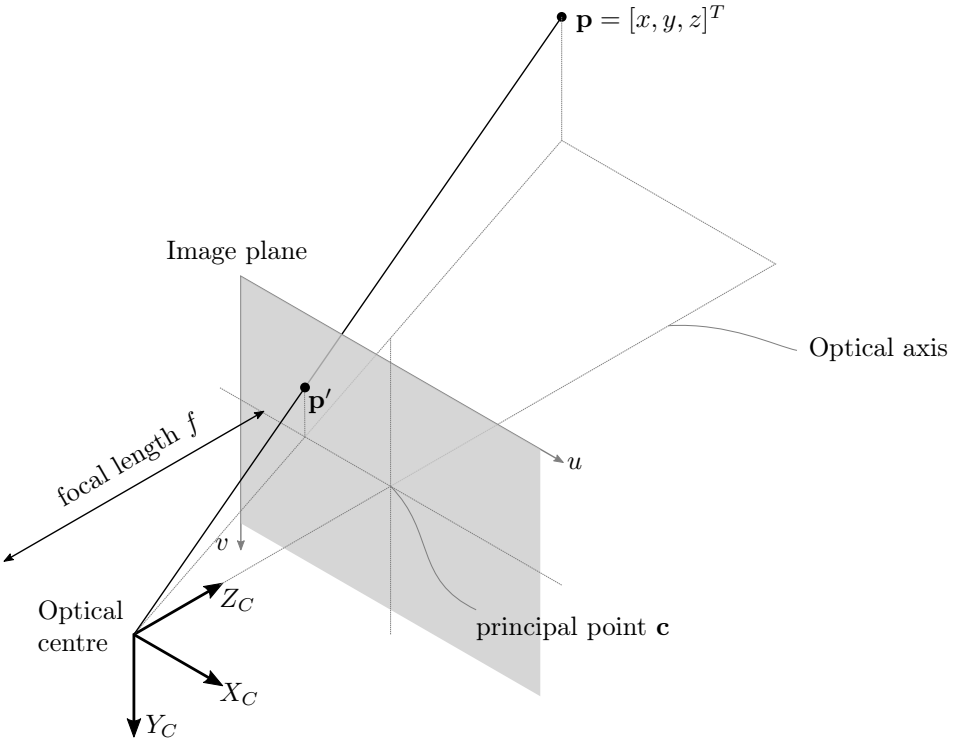


Figure 3.3: Projection in the pinhole camera model.

Image formation in the pinhole model is explained by assuming an infinitely small aperture, hence the term pinhole. Consider the illustration in Figure 3.3. All light rays converge on the optical centre, which coincides with the origin of the camera reference frame. The distance from the optical centre to the principal point $\mathbf{c}$ on the image plane is equal to the focal length f. A light ray from point $\mathbf{p} = [x, y, z]^T$ passes through the optical centre, thereby illuminating the point $\mathbf{p}' = [x', y', z']^T$ situated on the image plane. The principal of similar triangles dictate that $\mathbf{p}' = [fx/z, fy/z, f]^T$. A point $[x, y, z]^T$ is therefore mapped to the point $[fx/z, fy/z, f]^T$ on the image plane. Ignoring depth, this projection is given by a $\mathbb{R}^3$ to $\mathbb{R}^2$ mapping [54, p. 154], i.e.

$$[x, y, z]^T \rightarrow [fx/z, fy/z]^T. \tag{3.1}$$

Introducing homogeneous coordinates, Equation (3.1) may be rewritten in matrix form as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \frac{1}{z}\mathbf{K}\mathbf{p}, \tag{3.2}$$

with the camera calibration matrix

$$\mathbf{K} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.3}$$

Equation (3.2) provides the framework for the conversion of points in 3D space to image coordinates. Generally, the point $\mathbf{p}'$ on the image plane will be sought in terms of pixel coordinates. The focal length entries in the camera calibration matrix can be

appropriately scaled to induce a $\mathbb{R}^3$ metric to $\mathbb{R}^2$ pixel transformation. In practice, the camera calibration matrix defaults to focal lengths in pixels, and it includes parameters to account for non-idealities such as an offset principal point or an unequal number of pixels per unit length in the respective dimensions [54, p. 155–157].

Camera calibration methods aim at estimating the parameters that constitute the camera calibration matrix. These quantities are referred to as intrinsic parameters. The discussion will, however, not delve into detail with regard to camera calibration. The technique of Zhang [55] has become the preferred standard, and is therefore used in this work for intrinsic calibration purposes.

3.2.2 Stereo Geometry

When multiple cameras view the same scene, depth information can be extracted from the images. For the same to be achieved with a single camera requires knowledge of the scene being recorded.

Consider the idealised horizontal stereo configuration shown in Figure 3.4. In this scenario, the optical centres and image planes of the respective cameras are coplanar. The optical centres are also some distance b apart, which is referred to as the stereo baseline. Calculating the horizontal and vertical coordinates, x and y, relies on the same principles as before. However, the availability of depth means that they are now uniquely determined. The sought quantity is the z coordinate of point $\mathbf{p}$. Define the *disparity* as

$$d = u_L - u_R, \tag{3.4}$$

where u_L and u_R are the horizontal coordinates where point $\mathbf{p}$ projects on the respective image planes. Assuming that the origin of the coordinate system coincides with the left camera centre, then

$$\begin{aligned} u_L = f\frac{x}{z}, \quad u_R = f\frac{x-b}{z}, \\ \rightarrow d = f\frac{x}{z} - f\frac{x-b}{z}, \\ = \frac{fb}{z}, \\ \rightarrow z = \frac{fb}{d}. \end{aligned} \tag{3.5}$$

Equation (3.5) gives the relation between disparity and range for calibrated cameras in an ideal horizontal stereo configuration. The key quantity that is required for depth calculation is the disparity. Determining the disparity relies on finding corresponding pixels in the respective images. Correspondence methods generally rely on texture information to perform inter frame matching.

In reality, achieving perfect camera alignment is impossible by means of physical arrangement. A general alignment between two cameras is shown in Figure 3.5. A point $\mathbf{p}$ projects to the point $\mathbf{p}'$ on the left camera's image plane. The correspondence of this point on the right image plane is constrained to the horizontal line drawn across the plane. In fact, any point that projects onto the line drawn on the left camera plane is constrained to coincide with the corresponding line on the right image plane. These lines are known as conjugate epipolar lines, and they are used to guide stereo correspondence searches.

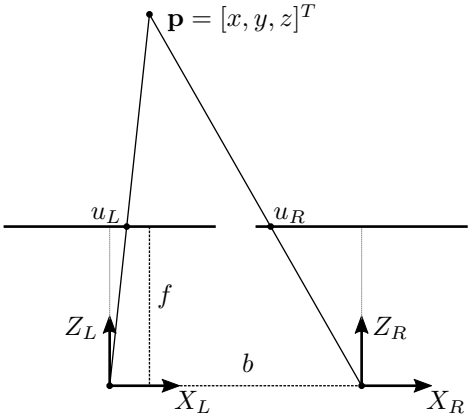


Figure 3.4: Ideal horizontal stereo vision geometry.

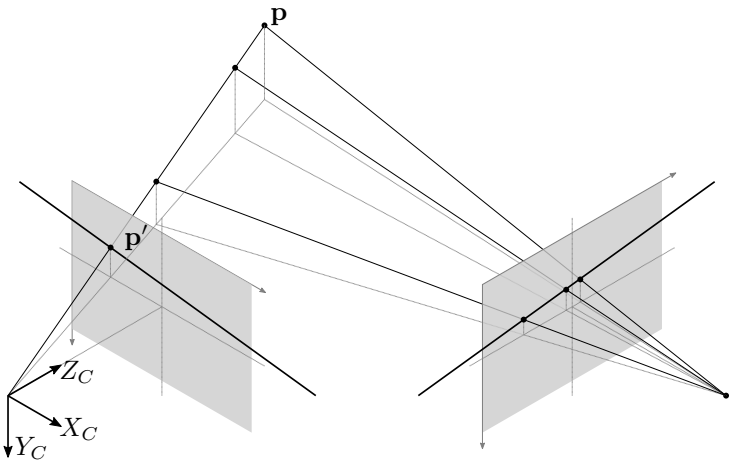


Figure 3.5: Epipolar geometry for general stereo camera alignments.

Accepted methods for the calculation of stereo correspondences assume ideal geometry as illustrated in Figure 3.4. Ideal geometry implies that distinctive pairs of epipolar lines are parallel, and that conjugate pairs are not vertically offset. This enables correspondence searches to simply be performed across the applicable row of stereo image pairs. Stereo images that have these ideal properties are termed *rectified*. The process of stereo rectification involves the transformation of both images by means of an image processing procedure. Rectification also requires knowledge of the physical arrangement between the two cameras, known as the extrinsic parameters. As with intrinsic camera calibration, standard methods exist for the calculation of extrinsic parameters for a stereo vision setup [56].

3.2.3 Phase Monopulse

Information from active radar sensors is processed quite differently than that of passive imaging sensors. The physical operating characteristics lead to measurements that are in a range-bearing format. Simple time-of-flight principles govern the calculation of range. As for bearing, traditional radar operation assume the antenna pointing direction to coincide with the detection angle. An inherent problem with this approach is that a narrow beam is required for accurate angle extraction, consequently hindering the ability to scan large volumes. Electronically scanned antenna arrays permit the best of both worlds due to

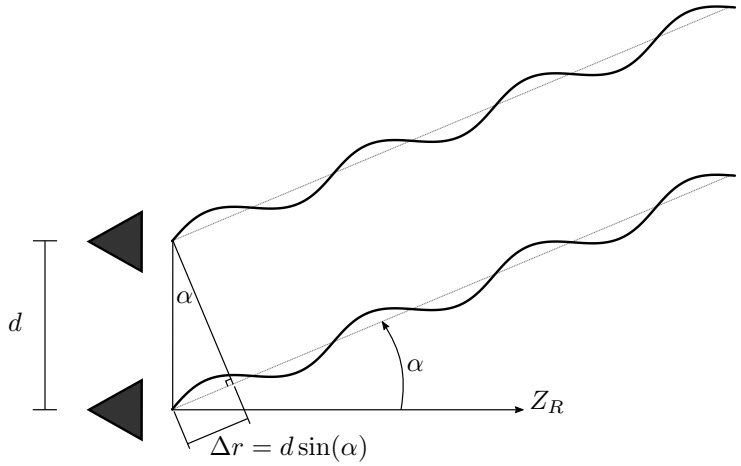


Figure 3.6: Illustration of the geometric principles that govern angle extraction by the principle of phase monopulse. The angled incidence of the incoming wavefront results in a phase difference at the respective receive antennas, which can be used to determine the angle of incidence.

rapid beam steering, but it requires intricate hardware solutions.

An alternative to beam steering is to use the information from multiple antennas to achieve within-beam angular discrimination by the principle of monopulse. In doing so, it is possible to maintain fine angular resolution, whilst simultaneously covering large swaths. Two techniques, namely amplitude and phase comparison monopulse, can be used, depending upon the antenna configuration. The antenna configuration of the radar used in this project permits the use of phase monopulse, since the respective antenna boresight axes are parallel [57, p. 165]. Section 4.2 will provide more details with regard to radar measurement extraction. The presentation here serves the purpose of discussing phase monopulse, since the next section describes a technique specific to such radars.

Figure 3.6 shows a top view illustration of a wavefront incident to the radar's receive antennas, which are drawn as black triangles. The wave is assumed to originate from a point scattered at an angle α from the boresight direction Z_R. Furthermore, the distance to the target, r, is assumed to be much greater than the separation of the antennas, known as the antenna baseline d. Due to the angled incidence, the distance that the wave travels to the respective antennas differs with a value of [57, p. 166]

$$\Delta r = d\sin(\alpha),$$

resulting in a phase difference of

$$\begin{aligned}\Delta\phi &= \frac{\Delta r}{\lambda}2\pi, \\ &= \frac{2\pi}{\lambda}d\sin(\alpha),\end{aligned} \tag{3.6}$$

where λ is the wavelength. The wavelength is a function of the chosen centre frequency, and the baseline can be measured or calibrated. Consequently, angle extraction requires only the phase of the two receive channels, which is available through the use of complex IQ receiver architectures. Note that the radar used here allows angle extraction in one dimension only. The available information is therefore two-dimensional (range and bearing).

3.3 Extrinsic Sensor Calibration

The first step in sensor fusion is the registration of measurements from the respective subsystems to a common frame of reference [30, p. 116]. An inaccurate estimate of the geometric offset between the sensors will cause faulty registrations, thereby impeding perception performance. The parameters that describe the alignment between sensors are known as extrinsic parameters. This section will describe the approach taken in this project to determine the extrinsic sensors-to-sensor parameters.

3.3.1 Problem Description

The problem at hand is to estimate the rigid body transformation between the two sensors' reference frames. The simplest way in which the extrinsic parameters describing this transformation may be acquired is by physically measuring the geometric arrangement. However, such a procedure may result in poor estimates due to the difficulty of accurately determining the actual origins of the respective sensors. Improved estimates should result from an automated calibration method, in which the sensors measure the same target. The calibration can then be cast into an optimisation problem in order to solve for the extrinsic parameters.

Suppose that by the process of Section 3.2.2, the camera takes a measurement of a point $\mathbf{p}^C$ in the CRF. The measurement, also in the CRF, is given by $\mathbf{z}^C_{\text{cam}}$. Moreover, using phase monopulse, the radar returns a measurement of the same object in the RRF as

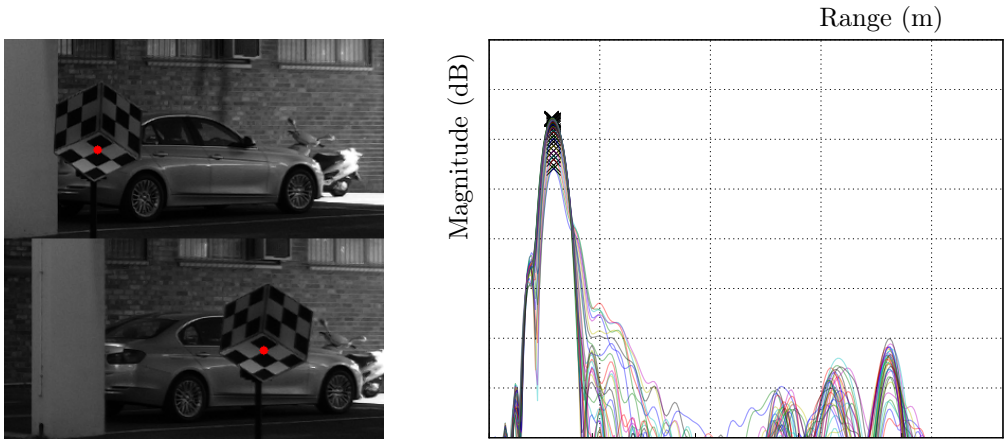
$$\mathbf{z}^R_{\text{rdr}} = [-r\sin(\alpha), 0, r\cos(\alpha)]^T, \tag{3.7}$$

where α is the counterclockwise angle from the boresight, i.e. the azimuth angle (see Figure 3.6). The notation $\mathbf{z}^R_{\text{rdr}}$ is used to describe a measurement $\mathbf{z}_{\text{rdr}}$ in the RRF. The same measurement's coordinate in the CRF is given by

$$\begin{aligned}\mathbf{z}^C_{\text{rdr}} &= [x^C_{\text{rdr}}, y^C_{\text{rdr}}, z^C_{\text{rdr}}]^T, \\ &= \mathbf{R}^C_R \mathbf{z}^R_{\text{rdr}} + \mathbf{t}^C_R,\end{aligned} \tag{3.8}$$

where $\mathbf{R}^C_R$ is the rotation matrix of the RRF with respect to the CRF, and $\mathbf{t}^C_R$ is the coordinate of the RRF's origin in the CRF. The three-dimensional rotation and translation described by Equation (3.8) follows the Euler-321 convention, in which the rotation is decomposed into a sequence of rotations about the z, y, and x axes of the CRF, given by the angles ψ, θ and, ϕ, respectively. The Euler parameterisation to three-dimensional transformations is fully defined by these angles and the elements of the translation vector $\mathbf{t}^C_R = [t_x, t_y, t_z]^T$. Determination of the extrinsic parameters therefore reduces to the solving of $\psi, \theta, \phi, t_x, t_y$, and t_z.

The discussion above assumes that the intrinsic parameters that impact the formation of the respective measurements $\mathbf{z}^C_{\text{cam}}$ and $\mathbf{z}^R_{\text{rdr}}$ are known. There is no need to adapt the standard methods that exist for the intrinsic and extrinsic calibration of stereo cameras. Therefore, the presentation will continue with the assumption that the stereo vision cameras are calibrated to produce rectified images. With regard to the radar, the antenna baseline may be viewed as an intrinsic parameter. If the baseline is specified incorrectly, extrinsic radar-to-camera parameters that result from the optimisation will be incorrect. Rather than measuring the baseline, it is included as an additional free parameters which is to be estimated in the calibration procedure.



(a) (b)

Figure 3.7: Manual measurement labelling for extrinsic calibration by means of (a) guided corner selection in stereo image pairs, and (b) guided local peak finding in radar range spectrums.

3.3.2 Measurement Extraction

A calibration target that is visible to both sensor subsystems is required for the calibration procedure. For this purpose, a simple corner reflector was enhanced by applying checkerboard paper patterns to the three reflecting surfaces facing the sensors (see Figure 3.7a). The resulting calibration target is both highly reflective (high RCS), and contains sharp corners that enable precise correspondences to be found in the stereo image pairs.

It is vital to assure that any measurement considered for parameter estimation originated from the calibration target. For this reason, the raw sensor data is manually labelled. Labelling radar data is as simple as identifying the peak in the range spectrum that corresponds to the calibration target. A rough estimate of the range to the target, along with the extremely high RCS of the corner reflector makes this process trivial. The extracted data consists of the range to the target as well as the signal's phase at both receivers. The phases are used in subsequent angle calculations. In an attempt to achieve increased accuracy, the average range and angle from 24 scans constitute a single radar calibration sample. The reason for choosing 24 pulses will become clear in Chapter 4. Figure 3.7b illustrates the local peaks that were found in the range spectrums of the respective pulses.

Image measurement extraction exploits the strong gradients of the checkerboard patterns to calculate disparity. A corner detection algorithm is manually guided to any of the corners on the calibration target in one of the images. This is repeated for the identical corner in the other image. An example pair of corners is shown in Figure 3.7a. The information is subsequently projected to the metric camera frame. The good features to track algorithm of Shi and Tomasi [58] is implemented for corner detection, while OpenCV's sub-pixel accuracy algorithm is used to refine the corners.

3.3.3 Parameter Estimation

It was mentioned earlier that the calibration problem is cast into an optimisation problem. The task of the optimisation procedure is to minimise an error function by varying the parameters that constitute the extrinsic calibration. With reference to the problem description given in Section 3.3.1, the measurements $\mathbf{z}^C_{\text{cam}}$ and $\mathbf{z}^C_{\text{rdr}}$ that originate from

the same point $\mathbf{p}^C$ should ideally be the same. Intuitively, an error function of the form $e = |\mathbf{z}^C_{\text{cam}} - \mathbf{z}^C_{\text{rdr}}|$ may seem applicable to minimise, where $|\cdot|$ denotes the norm operator. However, certain physical limitations suggest an alternative formulation. As mentioned in the beginning of the chapter, the radar's operating bandwidth is 155 MHz. The theoretical range resolution resulting from this quantity equates to a coarse 0.97 m[1]. As a consequence, incorporating extrinsic parameters relating to the offset in the optimisation may lead to inaccurate results, and it is therefore excluded. Instead of optimizing the full set of rotation and translation parameters, the optimisation is restricted to the two quantities that influence azimuth angle reporting, namely the antenna baseline d and the rotation θ about the Y axis. Moreover, the optimisation error function is based solely on the angular error, and is of the form

$$e = \left| \angle_\alpha \mathbf{z}^C_{\text{cam}} - \angle_\alpha \mathbf{z}^C_{\text{rdr}} \right|, \tag{3.9}$$

where $\angle_\alpha(\cdot) = \arctan(x/z)$ denotes the azimuth angle of the measurement $(\cdot)$.

The formal optimisation procedure is based on numerous observation pairs from the vision and radar subsystems. Each pair undergoes the manual measurement extraction procedure, before the optimisation is commenced. With the parameter dependence stated explicitly, the error function of Equation (3.9) for the i-th measurement pair is of the form

$$e_i(\theta, b) = \left| \angle_\alpha \mathbf{z}^C_{\text{cam},i} - \angle_\alpha \left(\mathbf{R}^C_R(\theta) \mathbf{z}^R_{\text{rdr},i}(b) + \mathbf{t}^C_R \right) \right|, \tag{3.10}$$

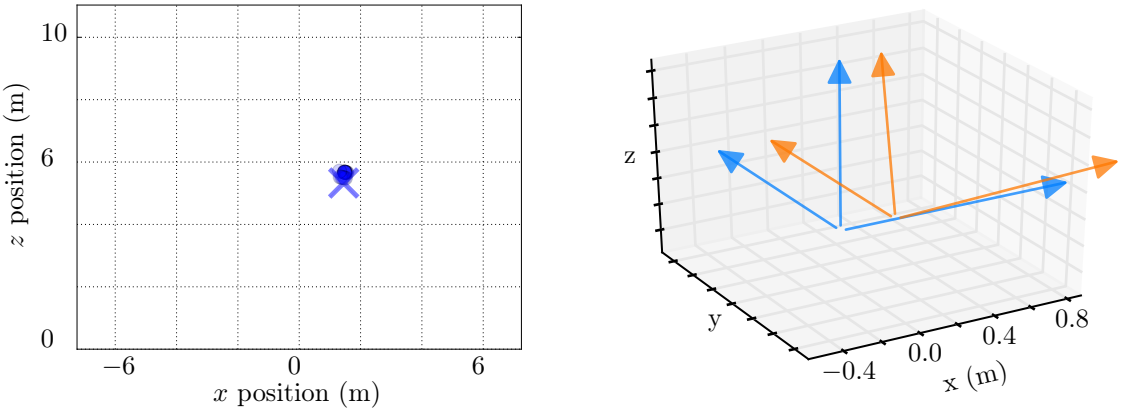
with $\mathbf{z}^R_{\text{rdr},i}(b)$ as in Equation (3.7). Given that N measurement pairs proceed to calibration, then, the complete error function for the set of measurements is

$$E(\theta, b) = \sum_{i=1}^{N} e_i(\theta, b). \tag{3.11}$$

The Levenberg-Marquardt algorithm [59, 60] is used to minimise this error function. An initial estimate of the rotation, translation and baseline distance is acquired from physical measurements, and passed to the algorithm. The parameters that are not allowed to vary remain at their measured values. Figure 3.8b shows the post calibration results of the position and rotation of the RRF in the CRF. The points from a sample measurement pair are plotted in Figure 3.8a, where the cross denotes the image measurement and the various circles represents respective radar measurements that have been projected to the camera frame using the optimisation solution.

The root-mean-square (RMS) reprojection error decreased from 0.56 m for the initial rotation and translation parameters to 0.42 m for the optimised parameters. Due to the imprecise nature of the measurements that result from the sensors, it would be impossible to acquire a perfect reprojection error. The error difference is fairly small, but indicates improved accuracy nevertheless. To the best of the author's knowledge, no existing technique address the camera-based intrinsic baseline calibration of a monopulse radar as described here. The proposed method is simple to implement, and delivers satisfactory results for the baseline distance as well as the axis rotation.

[1]The theoretical range resolution of a radar is inversely proportional to its bandwidth, and is given by $\delta r = \frac{c}{2B}$, where c is the propagation speed, and B is the bandwidth [12, p. 690].



(a) Vision (cross) and radar (circles) measurements in the CRF following extrinsic calibration.

(b) Rotation and translation of the RRF (orange) in the CRF (blue).

Figure 3.8: Extrinsic calibration results. (a) Reprojected points from both sensors following the calibration. (b) The relative axis alignment returned by the optimisation procedure.

Chapter 4

Measurement Extraction

The literature review chapter briefly described some of the notable measurement extraction approaches found among the common exteroceptive sensors. All DATMO methods require sensor measurements to undergo some form of processing, with the purpose of extracting valuable information that describes moving objects in the surrounding environment. Although the detection facet is not the primary focus of this project, as a practical grounding, the discussion proceeds to the presentation of the radar and vision measurement extraction techniques implemented in this work.

4.1 Vision

Spatio-temporal information acquired from a sparse feature tracking framework is used to identify moving image regions. Appearance information is wholly disregarded so to not yield object-specific results. Figure 8.4 shows the basic workflow of the measurement extraction algorithm, which can be summarised as follows: A sparse feature detector identifies strong candidate features which are subsequently tracked. The tracker's filtering induces smoothing, enables basic outlier removal, and provides a framework for trajectory storage and analysis. The available information is ultimately processed in a clustering routine which groups similar feature tracks.

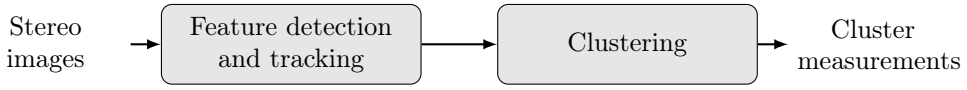


Figure 4.1: Diagram of the stereo vision measurement extraction algorithm.

4.1.1 Feature Detection

Feature detection is conducted using the features from accelerated segment test (FAST) algorithm of Rosten and Drummond [61]. The detector significantly outperforms other alternatives with regard to computational complexity, and is tailored for consistent multi-view feature extraction [61]. These attributes favour the use of the FAST corner detector for real-time stereo vision applications.

The use of a sparse feature detection carries some inherent disadvantages. The most notable of these is the inability to gather information from low-textured image regions,

which may impede the eventual ability to accurately estimate target extent. Extent information should predominantly be extracted from the vision subsystem, since it offers substantial higher resolution than the radar. In order to alleviate possible negative effects, feature detection thresholds are set to produce semi-dense information, i.e. thousands of features, spread across the field of view, are identified as candidates for tracking. By this approach, fairly accurate extent information is available, while the computation demands remain considerably lower than for dense detection methods.

4.1.2 Feature Tracking

The temporal data required for the functioning of the algorithm is available through a state estimator that tracks the detected features over time. Resulting motion information is of great value for moving object segmentation. The remainder of this section will detail the feature tracking framework implemented in this project.

Kalman Filter

The semi-dense requirements set out in Section 4.1.1 calls for careful consideration with regard to subsequent processing. Tracking features that may number in their thousands necessitates a very efficient state estimator. To this end, the Kalman filter is introduced. The Kalman filter is a realisable formulation of the Bayes filter recursive Equations (2.3) and (2.4). Instead of propagating the full target state density, a Gaussian approximation is adopted, i.e.

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \mathcal{N}(\mathbf{x}_k; \mathbf{m}_k, \mathbf{P}_k), \tag{4.1}$$

where $\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{P})$ denotes the Gaussian distribution defined over the vector $\mathbf{x}$ with mean $\mathbf{m}$ and covariance $\mathbf{P}$. An important constraint of the Kalman filter recursive equations is that the it must retain the Gaussian structure of the state distribution. This implies that the dynamic and measurement models must be linear Gaussian transformations. Note that the control input found in traditional control systems is not included in the prediction update, since this quantity is unknown in target tracking.

For linear models, the following equations define the Kalman filter's predict-correct recursion [9, p. 42]:

$$\mathbf{m}_{k|k-1} = \mathbf{F}_k \mathbf{m}_{k-1|k-1}, \tag{4.2}$$
$$\mathbf{P}_{k|k-1} = \mathbf{Q}_k + \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T, \tag{4.3}$$
$$\mathbf{S}_{k|k-1} = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k, \tag{4.4}$$
$$\mathbf{K}_{k|k-1} = \mathbf{P}_{k|k-1} \mathbf{H}_k \mathbf{S}_{k|k-1}^{-1}, \tag{4.5}$$
$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_{k|k-1}(\mathbf{z}_k - \mathbf{H}_k \mathbf{m}_{k|k-1}), \tag{4.6}$$
$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_{k|k-1} \mathbf{H}_k) \mathbf{P}_{k|k-1}, \tag{4.7}$$

where $\mathbf{F}_k$ is the state transition matrix, $\mathbf{Q}_k$ is the process noise covariance, $\mathbf{H}_k$ is the observation matrix, and $\mathbf{R}_k$ is the measurement noise covariance. Equations (4.2) and (4.3) define the prediction update, while Equations (4.4) to (4.7) define the measurement update using the associated measurement $\mathbf{z}_k$.

The prediction update is essentially a transformation of the state distribution through the linear dynamic model described by $\mathbf{F}_k$ and $\mathbf{Q}_k$. The transition matrix describes the deterministic relation between the state mean at time $k-1$ and the prior at time k, while the process noise covariance matrix models the uncertainty in the transition. In target

tracking, these matrices are derived from the target's motion model. The measurement update incorporates the observation to refine the state estimate. The key quantity here is the *innovation*, which is the difference between the expected measurement $\mathbf{H}_k\mathbf{m}_{k|k-1}$ and the actual measurement $\mathbf{z}_k$. The innovation covariance is given by the matrix $\mathbf{S}_{k|k-1}$. The innovation, together with its covariance, impact the resulting update in Equation (4.6) through means of the Kalman gain $\mathbf{K}_{k|k-1}$.

For a small state vector dimensionality, The Kalman filter provides a very efficient framework for Bayesian state estimation, as its recursive equations merely relies on simple matrix multiplications. Non-linear dynamic and measurement models can be incorporated by approximating the non-linear transformations as Gaussian. A Taylor series expansion and the unscented transform [62] characterise the extended and unscented Kalman filters respectively, the two of which being the most common non-linear extensions to the Kalman filter.

State Space Model

Measurements of image features are available in image coordinates, but target tracks in the context of DATMO are required in inertial coordinates. Tracking in the inertial space using image plane measurements would require non-linear approximation techniques such as those that have just been mentioned. The implementation of non-linear estimation methods would, however, significantly increase the tracker's computational demands. In a trade-off between optimality and speed, the latter is chosen for sparse feature tracking, i.e. features are tracked in image coordinates, which enable the use of the standard linear Kalman filter. The fact that image coordinates are not inertial coordinates may result in strange non-linear effects due to the accelerating coordinate frame. The linear Kalman filter is therefore an approximation.

Next, the dynamic and measurement models for the feature tracker are laid out. The motion of each feature point is modelled using linear dynamics with Gaussian noise according to the constant velocity model [63]

$$\mathbf{x} = [u, v_u]^T , \tag{4.8}$$

$$\mathbf{F} = \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q} = \sigma_w^2 \begin{bmatrix} \frac{1}{4}\Delta T^4 & \frac{1}{2}\Delta T^3 \\ \frac{1}{2}\Delta T^3 & \Delta T^2 \end{bmatrix}, \tag{4.9}$$

where u is the feature's horizontal image coordinate and v_u is its velocity, σ_w is the acceleration noise standard deviation, and ΔT is the time step. Exact decoupled trackers are implemented for the remaining axes that constitute an image coordinate, namely the vertical coordinate v, and the disparity d. It is therefore assumed that the respective image plane dimensions are wholly independent. Decoupled lower-order Kalman filters are more efficient due to the smaller dimensionality of the matrices that must be inverted.

Disparity information is incorporated through the use of OpenCV's stereo block matching dense stereo correspondence algorithm. Although a per-feature disparity search is expected to increase the efficiency, the additional research effort was not warranted. Here, disparity is needed only as a proof-of-concept. Consequently, features are extracted from the left image only, and the corresponding disparity is available from the dense correspondence algorithm.

The feature tracking measurement model is also linear and is of the form

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{4.10}$$

$$\mathbf{R} = \left[\sigma_u^2\right]. \tag{4.11}$$

where σ_u is the standard deviations of the measurement noise in the horizontal dimension. The same update matrices apply for the other dimensions.

An outliers rejection strategy based on the innovation is incorporated in the feature tracking framework. If the innovation is more than a constant factor standard deviations from the square root of the innovation covariance, i.e.

$$|(\mathbf{z}_k - \mathbf{H}_k\mathbf{m}_{k|k-1})| > c(\mathbf{S}_{k|k-1})^{1/2}, \tag{4.12}$$

where c is a constant, then the track is deleted. This rule is implemented in all three tracking dimensions, and if any fails the test, the track is removed. Note that the quantities in Equation (4.12) reduce to length 1 vectors and 1×1 matrices, i.e. they are scalars. In addition to outlier-based track deletion, M/N logic rules [44, p. 403] are also implemented for track management.

The consequence of choosing a linear image plane state space model is that acceleration effects may result from the fact that the underlying features' frame and the tracking frame are not the same. Although not ideal, the tracker induces track smoothing effects and allows for simple outlier removal, while the computational overhead is reduced to a minimum.

4.1.3 Data Association

No mention have yet been made with regard to the measurement-to-track association of the feature tracker. Implementing complex data association methods such as multiple hypothesis tracking will certainly be infeasible due to the sheer number of features being tracked. A different strategy is adopted that deviates from the traditional tracking approaches. Instead of regarding the output of the feature detector as measurements that are to be associated to existing track entities, new detections simply label a feature as a candidate for tracking. The appropriate measurement for any existing feature track is rather found by actively searching for its correspondence using optical flow principles. In particular, the pyramidal implementation of the Lucas-Kanade feature tracker [64] is used, and its output regarded as the 'measurement' of the feature. This measurement is then subsequently processed using the Kalman filter's measurement update equations. Disparity is acquired by querying the dense depth image. Resulting feature tracks provide velocity information for adequately textured image regions. The term for these velocities is *scene flow*, which is the extension of two-dimensional optical flow with depth information.

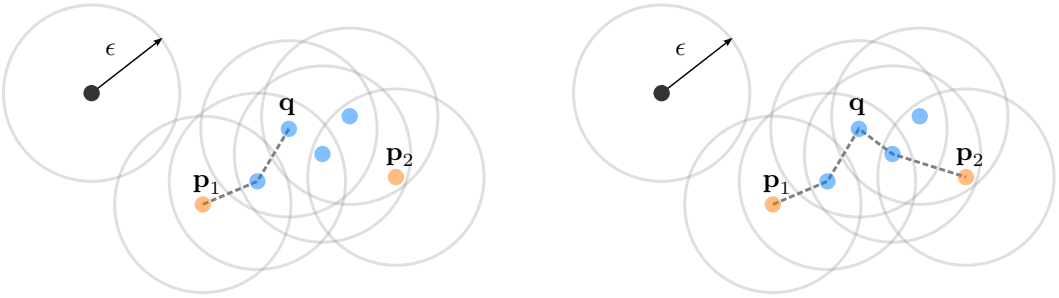
4.1.4 Clustering

To extract the eventual measurements from the vision subsystem, a clustering routine is implemented on the feature track data. The aim of clustering is to identify features tracks that originate from the same object. Ideally, minimum prior knowledge concerning the data should be required for the successful grouping of feature tracks. For instance, the number of targets that are present in any given scan is an unknown parameter that should be determined by the DATMO algorithm using exteroceptive sensor data. Many popular clustering algorithms such as k-means and expectation maximisation are unsuitable for this particular application, due to their dependence and sensitivity to the exact number of clusters. In this project, the chosen solution for feature track clustering with minimal prior

knowledge is the density-based spatial clustering of applications with noise (DBSCAN) algorithm [65]. DBSCAN identifies and groups dense areas in spatial datasets, whilst also providing an outlier labelling. DBSCAN requires only two parameters, namely the minimum number of samples a cluster may appropriate and a special distance threshold.

DBSCAN

The formalisation of what constitutes a cluster in the DBSCAN algorithm relies on a few definitions. Firstly, the ϵ-neighbourhood $N_\epsilon(\mathbf{p})$ of a point $\mathbf{p}$ is the set of all points that are a distance smaller than or equal than ϵ from $\mathbf{p}$. Points are labelled as core points if their ϵ-neighbourhood contains at least the minimum number of points. Any point $\mathbf{p}$ in a core point $\mathbf{q}$'s ϵ-neighbourhood is said to be directly density-reachable from the core point. Any point $\mathbf{p}$ is density-reachable from a core point $\mathbf{q}$ if there is a chain of points $\mathbf{p}_1, \ldots, \mathbf{p}_n$, $\mathbf{p}_1 = \mathbf{q}$, $\mathbf{p}_n = \mathbf{p}$ such that $\mathbf{p}_{i+1}$ is directly density-reachable from $\mathbf{p}_i$ [65]. A symmetric variant of the density-reachable condition is density-connected. Two points $\mathbf{p}$ and $\mathbf{q}$ are density-connected if their exists a point $\mathbf{o}$ from which both points are density-reachable. A cluster is defined as all points that are density-connected. Points that are density-reachable from any point within the cluster are included as well. Unreachable points are labelled as outliers. Figure 4.2 shows the different reachability definitions graphically.



(a) (b)

Figure 4.2: DBSCAN reachability illustration for a minimum samples parameter equal to 4 points. The blue circles are core points, since their ϵ-neighbourhood contains at least the minimum number of points. (a) Point $\mathbf{p}_1$ is density-reachable from $\mathbf{q}$, (b) while $\mathbf{p}_1$ and $\mathbf{p}_2$ are density-connected to each other by the core point $\mathbf{q}$. The black point is labelled an outlier.

DBSCAN clustering presents favourable characteristics in the sense that it requires very little assumptions about the dataset. The number of clusters does not have to be specified, and arbitrary shaped clusters may be found. Furthermore, outliers are inherently detected. A drawback of the algorithm is its inability to cluster datasets with varying degrees of densities.

To apply the DBSCAN clustering algorithm on the tracked points, the data is organised into a $N_{\mathrm{f},k} \times d$ clustering matrix $\mathbf{C}$, where $N_{\mathrm{f},k}$ is the number of tracked feature points at time k, and d is the number of dimensions included in the clustering. It is expected that tracks originating from a single object will demonstrate similar temporal behaviour, whilst being closely spaced in Euclidean distance. In view of this, each row of the clustering

matrix is a feature vector of the form

$$[(\mathbf{p}_i^W)^T, c(\mathbf{v}_i^W)^T]^T, \tag{4.13}$$

where $\mathbf{p}_i^W$ denotes the mean of the tracked point $\mathbf{p}_i$ in the world reference frame, and $\mathbf{v}_i$ is the average velocity of the point over a limited number of past time steps. The constant c is a weighting factor that scales the contribution of the velocity component. The implementation favours the velocity dimension with a 3:1 ratio, i.e. $c = 3$. Plain Euclidean distance is used as a dissimilarity metric during clustering. Clusters that result from the feature vector of Equation (4.13) are expected to contain points that exhibit similar spatial and temporal behaviour.

A labelling that indicates the make up of clusters and outliers result from the DBSCAN procedure. Outlier points are immediately discarded. Inlier clusters are thresholded on their average velocity in order to get the eventual measurements from the vision system. Each output clusters is expected to contain points that originate from the same moving object. In mathematical terms, a cluster measurement may be written in terms of a set of points, i.e. $Z = \{\mathbf{z}^{(i)}\}_{i=1}^{N_c}$, where N_c is the number of points in the particular cluster. Numerous clusters may be extracted at any given time step, suggesting a set of sets representation for the measurements extracted from the vision system, i.e. $Z_{\text{cam},k} = \{Z_k^{(i)}\}_{i=1}^{N_{\text{cam},k}}$, where $N_{\text{cam},k}$ is the number of cluster measurements at time k. Note that the form of the clustering feature vector of Equation (4.13) leads to output measurements that are in the WRF .

4.2 Radar

Radar-based measurement extraction is performed using standard two-dimensional Fourier analysis [12, p. 627]. Combined with the radar's continuous wave constraint, this translates to the frequency modulated continuous wave (FMCW) mode of operation [66]. The following presentation of FMCW will resemble content of Lipa and Barrick [67].

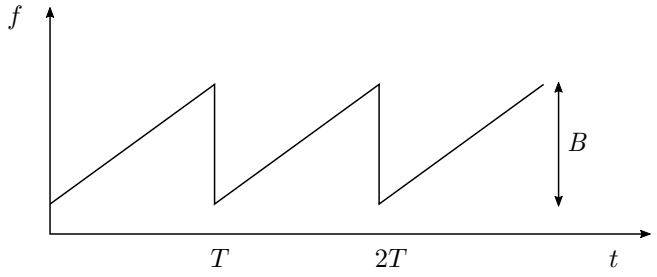


Figure 4.3: Linear frequency modulation waveform.

4.2.1 FMCW Radar Operation

The radar implements linear frequency modulation (LFM). For LFM, the frequency of the transmit waveform is of the form shown in Figure 4.3. The frequency at time t within each sweep is given by

$$f = f_c + \frac{B}{T}t, \tag{4.14}$$

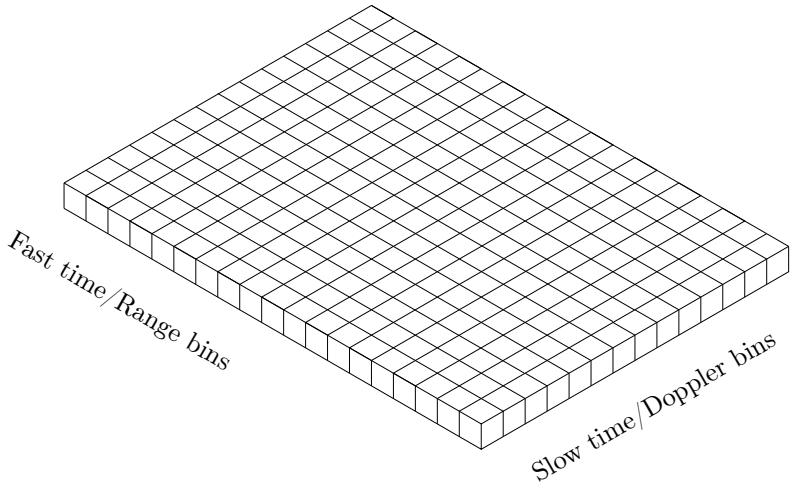


Figure 4.4: Two-dimensional slow-time fast-time pulse matrix.

where B is the sweep bandwidth, T is the pulse period, and f_c is the carrier frequency. An expression for the transmit signal v_{TX} can be obtained by integrating the frequency with respect to time, i.e.

$$v_{TX} = \cos\left(2\pi f_c t + \frac{\pi B t^2}{T}\right). \tag{4.15}$$

The received signal will be a delayed and attenuated version of the transmitted signal. Before digital processing, the received signal is mixed down to an intermediate frequency (IF). After some manipulation, and assuming that high frequency terms are filtered out, the IF signal for the n^{th} pulse is given by

$$v_{IF} = A\cos\ 2\pi\left(\frac{4\pi f_c \tau}{c} + \frac{2\pi B\tau(t-nT) - \pi B\tau^2}{T}\right), \tag{4.16}$$

where τ is the time-of-flight to the target and back, and c is the speed of propagation. It can be shown that the dominant frequency components in the mixed down IF signal are due to the time-of-flight and the Doppler shift induced by relative radial motion between the radar and the reflector. The former is directly related to the range r to the reflecting object. Lipa and Barrick [67] show that the resulting frequency from a reflector at a distance r is given by

$$f_p = \frac{2r}{c} \cdot \frac{B}{T}, \tag{4.17}$$

where the factor B/T is the chirp rate. The frequency that results from Doppler shifts is given by

$$f_d = \frac{2 f_c v_r}{c}, \tag{4.18}$$

where v_r is the relative radial velocity between the radar and the reflecting object.

Equations (4.17) and (4.18) enable the calculation of the range and relative radial velocity of a target. The standard detection method involves the use of Fourier frequency analysis to reveal the dominant frequency components of the received IF signal. In order to extract both range and velocity simultaneously, a pulse matrix as shown in Figure 4.4 is constructed. M samples are taken of the IF signal that correspond to a single frequency sweep of period T. These samples are then arranged into one column of the pulse matrix. This process is repeated for N consecutive sweeps to form the eventual $M \times N$ matrix.

The dimension along a single column (single sweep) is known as *fast-time*, whereas the inter-pulse dimension is known as *slow-time*. The time interval which constitutes the gathering of the $M \times N$ pulse matrix is known as a coherent processing interval (CPI). In this project, the radar detection algorithm used 128 fast-time samples and 24 repetitions in slow-time. The 24 sweeps that constitute a single CPI explain the use of 24 pulses in the calibration procedure of Section 3.3.

Two-dimensional Fourier analysis in the respective dimensions of the pulse matrix is implemented to extract target range and relative velocity. For this purpose, the fast Fourier transform (FFT) algorithm is utilised. The fast-time FFT is calculated for every pulse in the matrix. Even though the resulting frequencies contain superimposed time-of-flight and Doppler components, the latter is small in relation with the time-of-flight component and may be ignored [67]. Peaks in the fast-time spectrum therefore relate to range according to Equation (4.17). Upon calculating the fast-time FFTs, the second set of FFTs are taken row-wise, i.e. in slow-time. These provide the Doppler information, which related to the relative radial velocity according to Equation (4.18).

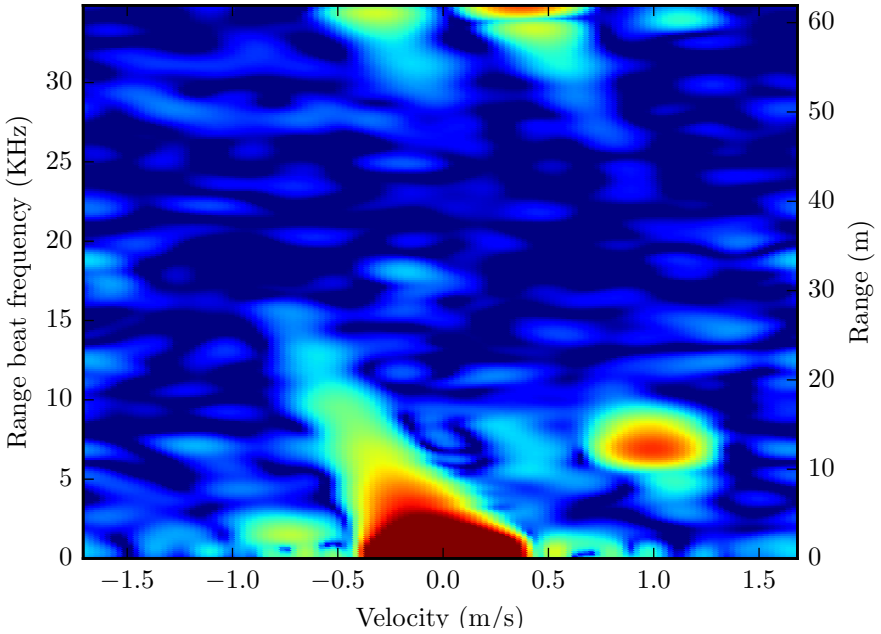


Figure 4.5: Complex range-Doppler map. The high energy region at approximately 12 m and centered at a velocity of 1 m/s corresponds to a vehicle.

Two-dimensional Fourier processing of a single CPI results in what is referred to as the *complex range-Doppler map* (CRDM). Figure 4.5 shows an example of the CRDM extracted whilst observing the front of a vehicle during highway driving. The local peak corresponding to the vehicle is visible as the prominent red blob at approximately 12 m and centered at a velocity of 1 m/s. Upon calculating the CRDM, the task is to identify areas in the spectrum with high energy content. Bin indices corresponding to such areas are subsequently treated as the measurements.

4.2.2 Constant False Alarm Rate Processing

Thresholding techniques are most often used for extracting measurements from CRDMs. Range propagation effects and clutter interference mean that a fixed threshold level will most definitely lead to poor detection performance. Consequently, standard radar measurement extraction methods rely on advanced thresholding algorithms aimed at achieving constant false alarm rates. The crux of constant false alarm rate (CFAR) processing is to accurately estimate the interference for all the cells in the CRDM. A CRDM bin's amplitude can then be tested against its interference to determine whether or not it should pass as a detection.

Different CFAR algorithms are characterised by the way in which the interference statistic is calculated. For radar measurement extraction in this project, the smallest-of cell-averaging constant false alarm rate (SOCA-CFAR) filter [68] is used, since it limits missed detections in the event that closely spaced targets induce masking effects [12, p. 612]. For any individual cell in the CRDM, referred to as a cell under test (CUT), the implementation calculates the average amplitude for a leading and lagging region, symmetrically offset in the range dimension, and chooses the smallest of the respective averages of the two regions as the interference statistic for the CUT. Given that σ_{SOCA} denotes the lower interference average for a CUT, then the detection threshold is

$$T_{\text{SOCA}} = c_{\text{CA}}\sigma_{\text{SOCA}}, \tag{4.19}$$

where c_{CA} is the CFAR constant. The CUT is marked as a detection if its amplitude exceeds the resulting threshold. Figure 4.6 illustrates the aforementioned SOCA-CFAR processing procedure.

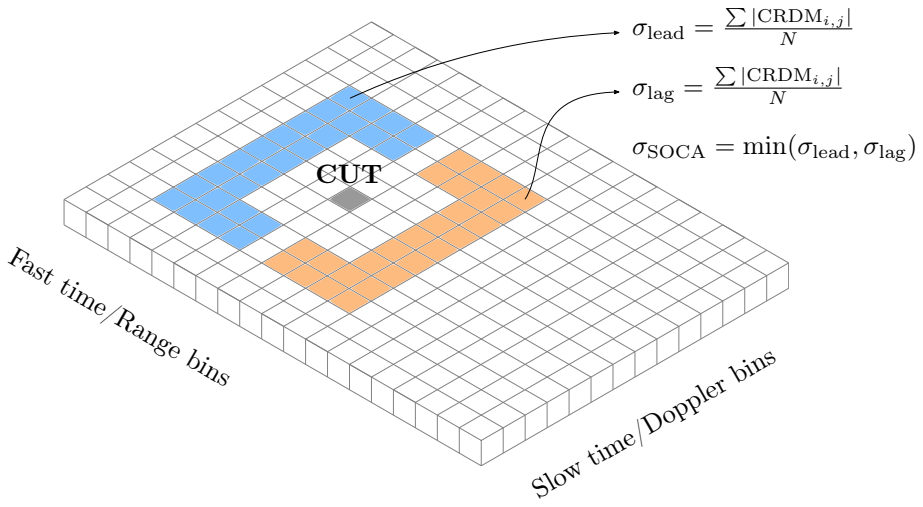


Figure 4.6: Calculation of the interference statistic for a single CRDM cell under test in the smallest-of cell-averaging constant false alarm rate algorithm. The result is an estimate of the interference for the particular CUT.

4.2.3 Measurement Post-Processing

CFAR processing results in bin indices that are labelled as detections, and from which range and radial velocity can be extracted. Additionally, the detection angle may also be calculated using the principle of phase monopulse discussed in Section 3.2.3. To determine

the azimuth angle, it is required to calculate the phase difference of the incoming signal at the respective receive antennas. It is desired to extract phase differences for the particular bins in the CRDM whose energies exceed the local interference. To do so, the method implemented by Molchanov et al. [69] is used, which calculates bin phase differences from the CRDMs of the two receive channels. The resulting angle for the bin (i, j) is given by

$$\alpha_{i,j} = \arcsin\left(\frac{\lambda(\angle\text{CRDM}_{i,j}^{\text{IF2}} - \angle\text{CRDM}_{i,j}^{\text{IF1}})}{2\pi d}\right), \tag{4.20}$$

where $\angle\text{CRDM}_{i,j}^{\text{IFn}}$ is the phase of bin (i, j) of the n^{th} receive IF channel. Only one channel's CRDM undergoes CFAR processing. Information from the 2^{nd} CRDM is used exclusively for angle calculations.

At the offset of angle extraction, each detection is in the range-bearing-velocity format $\mathbf{z} = [r, \alpha, v_r]^T$. To finalise the measurement extraction procedure, another routine of the DBSCAN clustering algorithm is implemented on the detection data in order to group nearby bins that correspond to the same target. The eventual measurements are given by the means of their clusters. Mathematically, a radar measurement set of $N_{\text{rdr},k}$ observations at time k is given by the set of vectors $Z_{\text{rdr},k} = \{\mathbf{z}_k^{(i)}\}_{i=1}^{N_{\text{rdr},k}}$, where the set elements are vectors of the form $\mathbf{z} = [r, \alpha]^T$.

The radial velocity is not included in the eventual measurement vectors, but is only used for the preceding clustering. The reason for excluding velocity from the measurement vector is due to the limited unambiguous velocity interval of the radar in the chosen mode of operation. Table 4.1 lists the unambiguous velocity interval along with other notable radar operating parameters. Qualitative analysis also revealed that the resolution of the radar is inadequate for use in an extended target measurement model, since the azimuth extent of measurement clusters exhibit a high degree of noise.

The methods laid out in this chapter address the detection facet of DATMO. The raw measurements that are extracted from the respective systems, however, are not suitable for direct use in higher level autonomous navigation applications. In the following chapters, techniques relating to state estimation and data fusion will be described, which should enable useful information to be inferred from the sensor measurements.

Table 4.1: Radar operating parameters.

Parameter	Symbol	Value
Sweep bandwidth	B	155 MHz
Sampling frequency	f_s	70 kHz
Fast-time samples	M	128
Slow-time samples	N	24
Pulse repetition frequency	PRF	$f_s/M = 547$ Hz
Maximum unambiguous range	r_{ua}	$cM/(4B) = 61.9$ m
Maximum unambiguous velocity	v_{ua}	$c/(4Tf_c) = 1.71$ m/s

CHAPTER 5

Multi-Target Tracking

The discussion at the end of literature review chapter outlined the areas where radar-vision fusion methods exhibit deficiencies. Advancing towards solutions, an in-depth analysis of multi-target tracking is required. To this end, the multi-target Bayes filter will be presented. The particular focus of this chapter will be on realistic realizations of the multi-target recursive Bayes filter, namely PHD filter variants. The PHD filter models both targets and observations as random finite sets (RFSs), thereby avoiding the complex problem of data association. RFS modelling resembles proper Bayesian multi-target tracking, and is therefore considered especially suited for the required task. The discussion will set off with a formal introduction to the multi-target Bayes filter. The remainder of the chapter will provide a detailed description of the *Gaussian mixture probability hypothesis density* (GM-PHD) filter [33], a realizable PHD variant.

5.1 Random Set Filtering

Consider the multi-target state finite set and the observation finite set of Section 2.3.4 at time k

$$X_k = \left\{ \mathbf{x}_k^{(i)} \right\}_{i=1}^{N_{\text{targets},k}}, \qquad \text{((2.5) revisited)}$$

$$Z_k = \left\{ \mathbf{z}_k^{(i)} \right\}_{i=1}^{N_{\text{measurements},k}}. \qquad \text{((2.6) revisited)}$$

In random set filtering, the uncertainty in these finite sets is modelled by random finite sets. An RFS $\mathcal{X}$ is simply a finite set valued random variable. The number of elements in an RFS are random and the elements themselves are also random, distinct and unordered [70]. The RFS cardinality (number of elements) is described by a discrete probability distribution, while, for a given cardinality, another density characterises the joint distribution of $\mathcal{X}$'s elements [33]. Before continuing, a mention of mathematical notation is in order.

- small letter boldface ($\mathbf{x}$) denotes a vector;
- calligraphic ($\mathcal{X}$) denotes a random finite set;
- capitalised italics (X) denotes a finite set; and
- capitalised boldface ($\mathbf{X}$) denotes a matrix.

5.1.1 Generic RFS Evolution Model

To get the discussion underway, the generic time evolution of the multi-target state RFS $\mathcal{X}$ will be described. The content is based on the presentation of Vo and Ma [33]. The evolution incorporates the following: target motion and death, target spawning, and spontaneous births as separate RFS models, the union of which addresses virtually all tracking scenarios. For a given multi-target state RFS $\mathcal{X}_{k-1}$ at time $k-1$, and conditioned on the existence at time k, target motion of each $\mathbf{x}_{k-1} \in \mathcal{X}_{k-1}$ is governed by the dynamic model $f(\mathbf{x}_k|\mathbf{x}_{k-1})$ describing the transition from time $k-1$ to k. This can be rewritten as an RFS generating functional $S_{k|k-1}(\mathbf{x}_{k-1})$ that can take on either $\{\mathbf{x}_k\}$ or $\emptyset$ depending on whether the target survives. Similar functionals can be constructed for the modelling of target spawning and births, leading to the prior state RFS

$$\mathcal{X}_{k|k-1} = \left[\bigcup_{\boldsymbol{\psi}\in\mathcal{X}_{k-1}} S_{k|k-1}(\boldsymbol{\psi})\right] \cup \left[\bigcup_{\boldsymbol{\psi}\in\mathcal{X}_{k-1}} B_{k|k-1}(\boldsymbol{\psi})\right] \cup \Gamma_k, \tag{5.1}$$

where $B_{k|k-1}(\boldsymbol{\psi})$ models targets spawned at time k from a target with previous state vector $\boldsymbol{\psi}$, and Γ_k models spontaneous births at time k. The transition model $S_{k|k-1}(\cdot)$ accounts for both the target dynamics and death. The latter is included as a state parameterised target survival probability $p_{S,k}(\boldsymbol{\psi})$. The spawn and birth models are problem dependent [33].

In the following paragraph, the RFS measurement model, which includes detection uncertainty and clutter as separate RFS models, will be described. For a given multi-target state RFS $\mathcal{X}_{k|k-1}$ at time k, and conditioned on detection at time k with probability $p_{D,k}(\mathbf{x}_{k|k-1})$, the expected observation $\mathbf{z}_k$ from each $\mathbf{x}_{k|k-1} \in \mathcal{X}_{k|k-1}$ is described by the measurement model $h(\mathbf{z}_k|\mathbf{x}_k)$. This process can again be rewritten as an RFS generating functional $\Theta_k(\mathbf{x}_k)$ that can take on either $\{\mathbf{z}_k\}$ or $\emptyset$ depending on whether the target is detected or not. The measurement set at time k is given by the union of target generated measurements and clutter, i.e.

$$\mathcal{Z}_k = K_k \cup \left[\bigcup_{\boldsymbol{\psi}\in\mathcal{X}_k} \Theta_k(\boldsymbol{\psi})\right], \tag{5.2}$$

where K_k models the clutter RFS at time k [33]. Equation (5.1) provides a generic RFS-based framework for the modelling of new targets and the fading of existing ones. A realisable example will be detailed in the following section. Equation (5.2) provides a comparable framework for the generic modelling of measurements.

5.1.2 Multi-target Bayes Filter

With the models describing the multi-target state RFS and observation RFS laid out, the focus shifts towards a recursive Bayesian formulation of multi-target inference. The goal is to estimate, at each time step, the current multi-target state set given all sensor measurements that have been collected, i.e. the posterior distribution $\mathcal{X}_{k|k} = f(X_k|Z_{1:k})$, where X_k is the finite set of target state vectors at time k, and $Z_{1:k}$ is the collection of finite measurement sets collected up to time k. The multi-target Bayes filter propagates the multi-target posterior via the recursive equations [32]

$$p(X_k|Z_{1:k-1}) = \int f(X_k|X_{k-1})p(X_{k-1}|Z_{1:k-1})dX_{k-1}, \tag{5.3}$$

$$p(X_k|Z_{1:k}) = \frac{h(Z_k|X_k)p(X_k|Z_{1:k-1})}{\int h(Z_k|X_k)p(X_k|Z_{1:k-1})dX_k}, \tag{5.4}$$

where

- X is the finite set of target state vectors;
- Z the finite set of measurements;
- $f(X_k|X)$ the multi-target dynamic model;
- $h(Z_k|X_k)$ the multi-target measurement model;
- $p(X_k|Z_{1:k-1})$ is the multi-target prior distribution; and
- $p(X_k|Z_{1:k})$ is the multi-target posterior distribution.

The reader will notice that the multi-target Bayes filter defined in Equations (5.3) and (5.4) resembles its single-target equivalent. Explicit expressions for the multi-target dynamic and measurement models can be derived using finite set statistics. They, however, are not required for the GM-PHD presentation [33], which is the subject of the next section.

5.2 Gaussian Mixture Probability Hypothesis Density Filter

The GM-PHD filter is a solution to the multi-target recursive Bayes filter in the same manner that the Kalman filter is a solution to the single-target recursive Bayes filter. Its tractability stems from a Gaussian approximation of the multi-target posterior distribution, in addition to assumed Gaussian dynamic and measurement models [33]. In contrast to the original PHD filter derived by Mahler [32], the Gaussian mixture implementation enables the explicit incorporation of state and measurement uncertainty models, whilst maintaining low complexity.

The key to realizing the multi-target recursion is the propagation of a lower order statistical moment of the posterior multi-target state set, termed the posterior *intensity* or probability hypothesis density. The intensity $v(\cdot)$ is a function on the multi-target state space, and the evaluation thereof results in a moment of the multi-target state RFS [33]. For the GM-PHD filter, the intensity is a Gaussian mixture[1]. To present the recursion of the intensity function, recall the evolution and measurement models presented in the previous section. Rather than modelling the full distributions, intensity functions akin to the intensity $v(\cdot)$ are required. The following notation defines the respective intensity functions:

$\beta(\cdot\|\boldsymbol{\psi})$	intensity of RFS $B(\cdot\|\boldsymbol{\psi})$ spawned by a target with state $\boldsymbol{\psi}$;
$\gamma(\cdot)$	intensity of the spontaneous birth RFS Γ; and
$\kappa(\cdot)$	intensity of clutter RFS K.

With the above spawn, birth and clutter definitions, it can be shown that the posterior intensity can be propagated in time via the recursive equations [33]

[1] A mixture is a weighted sum of probability distributions [71]. The weights are positive and are not required to sum to one. Individual distributions within a mixture are called *components*.

$$\upsilon(\mathbf{x}_k|Z_{1:k-1}) = \int p_S(\mathbf{x}_{k-1}) f(\mathbf{x}_k|\mathbf{x}_{k-1}) \upsilon(\mathbf{x}_{k-1}|Z_{1:k-1}) d\mathbf{x}_{k-1} + \int \beta(\mathbf{x}_k|\mathbf{x}_{k-1}) \upsilon(\mathbf{x}_{k-1}|Z_{1:k-1}) d\mathbf{x}_{k-1} + \gamma(\mathbf{x}_k), \quad (5.5)$$

$$\upsilon(\mathbf{x}_k|Z_{1:k}) = [1 - p_D(\mathbf{x}_k)]\upsilon(\mathbf{x}_k|Z_{1:k-1}) + \sum_{\mathbf{z}\in Z_k} \frac{p_D(\mathbf{x}_k) h(\mathbf{z}|\mathbf{x}_k) \upsilon(\mathbf{x}_k|Z_{1:k-1})}{\kappa(\mathbf{z}) + \int p_D(\mathbf{x}_k) h(\mathbf{z}|\mathbf{x}_k) \upsilon(\mathbf{x}_k|Z_{1:k-1}) d\mathbf{x}_k}. \quad (5.6)$$

Equations (5.5) and (5.6) describe the general form for the propagation of lower order statistical moments of the multi-target state RFS, and includes the transition, birth, death, spawn, clutter and measurement models.

Vo and Ma [33] proved that a closed-form solution for the above PHD recursion exists for a Gaussian mixture approximation of the posterior state RFS. The Gaussian mixture posterior is in fact a result of assumptions with regards to the birth and spawn intensity functions. Following the derivation of Vo and Ma [33], these, and other required assumptions are detailed below. The exhaustive list may be found in the original work of Vo and Ma [33].

A.1: Each target follows a linear Gaussian dynamic model and the measurement model is also a linear Gaussian, i.e.

$$f(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k; \mathbf{F}_k \mathbf{m}_{k-1|k-1}, \mathbf{Q}_k), \quad (5.7)$$
$$h(\mathbf{z}_k|\mathbf{x}_k) = \mathcal{N}(\mathbf{z}_k; \mathbf{H}_k \mathbf{m}_{k|k-1}, \mathbf{R}_k), \quad (5.8)$$

where $\mathbf{F}_k$ is the state transition matrix, $\mathbf{Q}_k$ is the process noise, $\mathbf{H}_k$ is the observation matrix, and $\mathbf{R}_k$ is the measurement noise.

A.2: The survival and detection probabilities are state independent, i.e.

$$p_S(\mathbf{x}) = p_S, \quad (5.9)$$
$$p_D(\mathbf{x}) = p_D. \quad (5.10)$$

A.3: The intensities of the spawn and spontaneous birth RFSs are Gaussian mixtures of the form

$$\beta(\mathbf{x}_k|\mathbf{x}_{k-1}) = \sum_{i=1}^{J_{\beta,k}} w_{\beta,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{F}_{\beta,k}^{(i)} \mathbf{m}_{k-1|k-1} + \mathbf{d}_{\beta,k}^{(i)}, \mathbf{Q}_{\beta,k}^{(i)}), \quad (5.11)$$

$$\gamma(\mathbf{x}_k) = \sum_{i=1}^{J_{\gamma,k}} w_{\gamma,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{\gamma,k}^{(i)}, \mathbf{P}_{\gamma,k}^{(i)}), \quad (5.12)$$

where $J_{\beta,k}$ is the number of spawn mixture components at time k, $w_{\beta,k}^{(i)}$ is the weight of each component, and $\mathbf{F}_{\beta,k}^{(i)}, \mathbf{d}_{\beta,k}^{(i)}, \mathbf{Q}_{\beta,k}^{(i)}$ describe the individual Gaussian distributions that are spawned off of a target state vector $\mathbf{x}_{k-1}$ whose mean is $\mathbf{m}_{k-1|k-1}$. Similarly, $J_{\gamma,k}$ is the number of spontaneous birth components, with mixture parameters $w_{\gamma,k}^{(i)}, \mathbf{m}_{\gamma,k}^{(i)}$ and $\mathbf{P}_{\gamma,k}^{(i)}$. Recall that the notation $\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{P})$ is used to indicate a Gaussian distribution defined over the vector $\mathbf{x}$ with a mean $\mathbf{m}$ and covariance $\mathbf{P}$. A few remarks may prove helpful in understanding the intensities. The mean vectors $\mathbf{F}_{\beta,k}^{(i)} \mathbf{m}_{k-1|k-1} + \mathbf{d}_{\beta,k}^{(i)}$ and $\mathbf{m}_{\gamma,k}^{(i)}$ are the respective peaks in the spawn and birth intensities, whilst their corresponding covariances

determine the uncertainty in the peak vicinities. The weight $w^{(i)}_{(\cdot),k}$ gives the expected number of targets described by the corresponding mixture component. The spawn intensity determines the behaviour of new targets that spawn from an existing target with previous state $\mathbf{x}_{k-1}$. The birth intensity determines where new targets originated from independently.

With the preliminaries having been explained, the discussion proceeds to the closed-form expressions of the GM-PHD recursion given by Equations (5.5) and (5.6). The derivation is governed by the aforementioned assumptions, and the presumption that the posterior intensity at time $k-1$ is a Gaussian mixture of the form

$$\upsilon(\mathbf{x}_{k-1}|Z_{1:k-1}) = \sum_{i=1}^{J_{k-1|k-1}} w^{(i)}_{k-1|k-1}\mathcal{N}(\mathbf{x}_{k-1};\mathbf{m}^{(i)}_{k-1|k-1},\mathbf{P}^{(i)}_{k-1|k-1}), \tag{5.13}$$

where $J_{k-1|k-1}$ is the number of posterior mixture components at time $k-1$. The first step of the recursion is the prediction of the intensity in Equation (5.13) to the next time step. The resulting intensity is given by

$$\upsilon(\mathbf{x}_k|Z_{1:k-1}) = \upsilon_S(\mathbf{x}_k|Z_{1:k-1}) + \upsilon_\beta(\mathbf{x}_k|Z_{1:k-1}) + \upsilon_\gamma(\mathbf{x}_k), \tag{5.14}$$

where the Gaussian mixture

$$\upsilon_S(\mathbf{x}_k|Z_{1:k-1}) = p_S \sum_{i=1}^{J_{k-1|k-1}} w^{(i)}_{k-1|k-1}\mathcal{N}(\mathbf{x}_k;\mathbf{m}^{(i)}_{S,k|k-1},\mathbf{P}^{(i)}_{S,k|k-1}), \tag{5.15}$$

with

$$\mathbf{m}^{(i)}_{S,k|k-1} = \mathbf{F}_k\mathbf{m}^{(i)}_{k-1|k-1}, \tag{5.16}$$

$$\mathbf{P}^{(i)}_{S,k|k-1} = \mathbf{Q}_k + \mathbf{F}_k\mathbf{P}^{(i)}_{k-1|k-1}\mathbf{F}_k^T, \tag{5.17}$$

models surviving targets and

$$\upsilon_\beta(\mathbf{x}_k|Z_{1:k-1}) = \sum_{i=1}^{J_{k-1|k-1}}\sum_{j=1}^{J_{\beta,k}} w^{(i)}_{k-1|k-1}w^{(j)}_{\beta,k|k-1}\mathcal{N}(\mathbf{x}_k;\mathbf{m}^{(i,j)}_{\beta,k|k-1},\mathbf{P}^{(i,j)}_{\beta,k|k-1}). \tag{5.18}$$

with

$$\mathbf{m}^{(i,j)}_{\beta,k|k-1} = \mathbf{F}^{(j)}_{\beta,k}\mathbf{m}^{(i)}_{k-1|k-1} + \mathbf{d}^{(j)}_{\beta,k}, \tag{5.19}$$

$$\mathbf{P}^{(i,j)}_{\beta,k|k-1} = \mathbf{Q}^{(j)}_{\beta,k} + \mathbf{F}^{(j)}_{\beta,k}\mathbf{P}^{(i)}_{k-1|k-1}(\mathbf{F}^{(j)}_{\beta,k})^T, \tag{5.20}$$

models targets that spawn from existing ones. The spontaneous birth intensity $\upsilon_\gamma(\mathbf{x}_k)$ is given by Equation (5.12). Equations (5.16) and (5.17) are the standard Kalman filter prediction equations, and Equation (5.15) implies that all surviving mixture components proceed to the prediction step. Upon combining the respective intensities, the predicted intensity of Equation (5.14) remains a Gaussian mixture and is of the form

$$\upsilon(\mathbf{x}_k|Z_{1:k-1}) = \sum_{i=1}^{J_{k|k-1}} w^{(i)}_{k|k-1}\mathcal{N}(\mathbf{x}_k;\mathbf{m}^{(i)}_{k|k-1},\mathbf{P}^{(i)}_{k|k-1}). \tag{5.21}$$

Supposing all the assumptions hold, then the posterior intensity at time k is also a Gaussian mixture and is given by

$$\upsilon(\mathbf{x}_k|Z_{1:k}) = (1-p_D)\upsilon(\mathbf{x}_k|Z_{1:k-1}) + \sum_{\mathbf{z}\in Z_k} \upsilon_D(\mathbf{x}_k;\mathbf{z}), \tag{5.22}$$

where Z_k is the finite set of measurements at time k, and

$$\upsilon_D(\mathbf{x}_k;\mathbf{z}) = \sum_{i=1}^{J_{k|k-1}} w_{k|k}^{(i)}(\mathbf{z})\mathcal{N}(\mathbf{x}_k;\mathbf{m}_{k|k}^{(i)}(\mathbf{z}),\mathbf{P}_{k|k}^{(i)}), \tag{5.23}$$

with

$$w_{k|k}^{(i)}(\mathbf{z}) = \frac{p_D w_{k|k-1}^{(i)} q_k^{(i)}(\mathbf{z})}{\kappa_k(\mathbf{z}) + p_D \sum_{j=1}^{J_{k|k-1}} w_{k|k-1}^{(j)} q_k^{(j)}(\mathbf{z})}, \tag{5.24}$$

$$\mathbf{m}_{k|k}^{(i)}(\mathbf{z}) = \mathbf{m}_{k|k-1}^{(i)} + \mathbf{K}_{k|k-1}^{(i)}(\mathbf{z} - \mathbf{H}_k\mathbf{m}_{k|k-1}^{(i)}), \tag{5.25}$$

$$\mathbf{P}_{k|k}^{(i)} = (\mathbf{I} - \mathbf{K}_{k|k-1}^{(i)}\mathbf{H}_k)\mathbf{P}_{k|k-1}^{(i)}, \tag{5.26}$$

$$\mathbf{K}_{k|k-1}^{(i)} = \mathbf{P}_{k|k-1}^{(i)}\mathbf{H}_k(\mathbf{S}_{k|k-1}^{(i)})^{-1} \tag{5.27}$$

$$\mathbf{S}_{k|k-1}^{(i)} = \mathbf{H}_k\mathbf{P}_{k|k-1}^{(i)}\mathbf{H}_k^T + \mathbf{R}_k. \tag{5.28}$$

The first term in Equation (5.22) explains missed detections, whereas the second represents the mixture components that result from new measurements. Equations (5.25) to (5.28) resemble the standard Kalman filter measurement update equations, and Equation (5.23) implies that all components in the prior mixture $\upsilon(\mathbf{x}_k|Z_{1:k-1})$ proceed to the update step with measurement $\mathbf{z}$. Hence, no data association is included. The weight assigned to new components, given by Equation (5.24), plays a key role in the facilitation of reliable tracking in the absence of data association: The factor $q_k^{(j)}(\mathbf{z})$ is the evaluation of the measurement likelihood function

$$q_k^{(j)}(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{H}_k\mathbf{m}_{k|k-1}^{(i)},\mathbf{S}_{k|k-1}^{(i)}) \tag{5.29}$$

at the position of measurement $\mathbf{z}$, where $\mathbf{m}_{k|k-1}^{(i)}$ and $\mathbf{P}_{k|k-1}^{(i)}$ are the prior state distribution parameters of the mixture component involved. Measurements that are distant from a component are unlikely, and will consequently give rise to low-weight components in the posterior intensity. A subsequent discussion will address the impact of component weights on track formation. Extracting target states from the Gaussian mixture posterior intensity $\upsilon(\mathbf{x}_k|Z_{1:k})$ is straightforward, since the means of the respective components comprise the local maxima of $\upsilon(\mathbf{x}_k|Z_{1:k})$. Furthermore, an estimate of the number of targets is available as the sum of the component weights.

The assumptions that are required to hold for the implementation of the filter are standard to most tracking applications [33], with exception of the linear dynamic and measurement models. However, the GM-PHD recursion may be altered to incorporate non-linear dynamic and measurement models in a similar vein in which the extended and unscented Kalman filters allow their incorporation into the standard Kalman filter [33].

Equations (5.14) and (5.22) are the high-level representation of the GM-PHD recursion. The framework allows for target spawning, spontaneous births and clutter measurements, as well as a Gaussian representation of uncertainty. The filter provides a realisable

probabilistic framework for multi-target tracking in the presence of missed detections and false alarms, whilst alleviating the need for explicit formulations of measurement-to-target correspondences.

CHAPTER 6

Extended Target Tracking

The foregoing chapter presented a rigorous Bayesian framework for multi-target tracking, but included no mention of target modelling. Standard modelling would entail the point target assumption, which is often invalidated in DATMO environments. The discussion in Section 2.6 raised the issue of naive extent modelling approaches which is prevalent in radar-vision literature. The subject of this chapter is extended target tracking in a Bayesian sense, i.e. joint kinematic and extent estimation. In this regard, the random matrix model will be presented, which adopts an elliptical representation of object shape. The use of random matrices in measurement models allow for proper Bayesian extended target tracking without drastic implications with regards to efficiency. These attributes motivate the use of random matrices for extended target modelling, particularly in practical applications. Note that the theory with regards to random matrix modelling is based on work of Koch [39] and Feldmann et al. [41]. The inclusion of the random matrix model into the GM-PHD filter will be the subsequent focus.

6.1 Random Matrix Modelling

The definition of an extended target states that such a target may give rise to numerous measurements per scan time. The appropriate mathematical representation for such measurements at time k is the collection of the individual point measurements, given by the set

$$Z_k = \left\{ \mathbf{z}_k^{(i)} \right\}_{i=1}^{N_k}, \tag{6.1}$$

where N_k is the number of measurements at time k. Equation (6.1) is essentially the same as the observation set of Equation (2.6). What is important to note here, however, is that Equation (6.1) describes a set of measurements that originate from a single object, whereas Equation (2.6) describes measurements that may originate from numerous objects or clutter. An extended target measurement set of the form given by Equation (6.1) will also be referred to as a *cluster measurement*.

To begin with, some preliminary definitions are needed. Firstly, the mean vector and measurement spread matrix of a cluster measurement are given by

$$\bar{Z}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{z}_k^{(i)}, \tag{6.2}$$

$$\tilde{Z}_k = \sum_{i=1}^{N_k} \left(\mathbf{z}_k^{(i)} - \bar{Z}_k\right)\left(\mathbf{z}_k^{(i)} - \bar{Z}_k\right)^T, \tag{6.3}$$

respectively. The mathematical development that is to follow involves explicit mention of kinematic and extent parameters and distributions. To distinguish these, target kinematic and extent distributions will be represented by the symbols $\mathbf{x}$ and $\mathbf{X}$ respectively. The random matrix approach to extended target tracking uses symmetric positive definite matrices to represent an object's extent, hence the use of uppercase notation.

6.1.1 Bayesian Formulation of Random Matrix Extended Target Tracking

The goal of the random matrix approach to extended target tracking is to infer the joint posterior density

$$p(\mathbf{x}_k, \mathbf{X}_k | Z_{1:k}) = p(\mathbf{x}_k | \mathbf{X}_k, Z_{1:k}) p(\mathbf{X}_k | Z_{1:k}), \tag{6.4}$$

where $\mathbf{x}_k$ is the centre point kinematic state vector at time k, $\mathbf{X}_k$ is a symmetric positive definite (SPD) matrix describing the target's extent at time k, and Z_k is an associated cluster measurement at time k [39]. The joint density in Equation (6.4) includes the factorisation into a vector-variate distribution $p(\mathbf{x}_k|\mathbf{X}_k, Z_{1:k})$ and a matrix-variate distribution $p(\mathbf{X}_k|Z_{1:k})$. The notation shows the explicit dependency of the kinematic distribution on the current object extent, which is what defines joint kinematic-extent estimation.

In a Bayesian framework, the estimation of the posterior distribution must be cast in a familiar recursive predict-correct progression. For the joint distribution of Equation (6.4), the prediction step is given by

$$p(\mathbf{x}_k, \mathbf{X}_k | Z_{1:k-1}) = \iint f(\mathbf{x}_k, \mathbf{X}_k | \mathbf{x}_{k-1}, \mathbf{X}_{k-1}) p(\mathbf{x}_{k-1}, \mathbf{X}_{k-1} | Z_{1:k-1}) d\mathbf{x}\, d\mathbf{X}, \tag{6.5}$$

where the simplifying Markov assumptions have already been included. The derivation in the original work of Koch [39] relies on the – often justifiable – assumption that the target's kinematic properties have no influence on the temporal evolution of its extent, conditioned on $\mathbf{X}_{k-1}$. The first factor of Equation (6.5), which is the dynamic model, may then be rewritten as

$$f(\mathbf{x}_k, \mathbf{X}_k | \mathbf{x}_{k-1}, \mathbf{X}_{k-1}) = f(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{X}_k) f(\mathbf{X}_k | \mathbf{X}_{k-1}). \tag{6.6}$$

The dependence of the kinematic evolution on extent remains intact, and cannot be ignored [39]. By combining Equations (6.4) to (6.6), the prediction step is given by

$$p(\mathbf{x}_k, \mathbf{X}_k | Z_{1:k-1}) = \iint f(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{X}_k) f(\mathbf{X}_k | \mathbf{X}_{k-1}) \\ p(\mathbf{x}_{k-1} | \mathbf{X}_{k-1}, Z_{1:k-1}) p(\mathbf{X}_{k-1} | Z_{1:k-1}) d\mathbf{x}\, d\mathbf{X}. \tag{6.7}$$

Further assuming that the temporal change of the object extent does not influence the prediction of the kinematic properties, i.e. $p(\mathbf{x}_{k-1}|\mathbf{X}_{k-1}, Z_{1:k-1}) = p(\mathbf{x}_{k-1}|\mathbf{X}_k, Z_{1:k-1})$, allows Equation (6.7) to be factored into independent integrations:

$$p(\mathbf{x}_k | \mathbf{X}_k, Z_{1:k-1}) = \int f(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{X}_k) p(\mathbf{x}_{k-1} | \mathbf{X}_k, Z_{1:k-1}) d\mathbf{x}, \tag{6.8}$$

$$p(\mathbf{X}_k|Z_{1:k-1}) = \int f(\mathbf{X}_k|\mathbf{X}_{k-1})p(\mathbf{X}_{k-1}|Z_{1:k-1})d\mathbf{X}. \tag{6.9}$$

The measurement update follows the prediction step. Application of Bayes' rule gives

$$p(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k}) = \frac{h(Z_k, N_k|\mathbf{x}_k, \mathbf{X}_k)p(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1})}{\int h(Z_k, N_k|\mathbf{x}_k, \mathbf{X}_k)p(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1})d\mathbf{x}\ d\mathbf{X}}, \tag{6.10}$$

where

$$h(Z_k, N_k|\mathbf{x}_k, \mathbf{X}_k) = h(Z_k|N_k, \mathbf{x}_k, \mathbf{X}_k)h(N_k|\mathbf{x}_k, \mathbf{X}_k) \tag{6.11}$$

is the factored form of the measurement model. The measurement model $h(Z_k, N_k|\mathbf{x}_k, \mathbf{X}_k)$ describes the distribution of measurements given the predicted target kinematic and extent parameters.

6.1.2 Gaussian Inverse Wishart Implementation

The following text gives a concise description of the recursion procedure derived by Koch [39], without delving into mathematical details. This should provide the required intuitive understanding for the subsequent presentation of Feldmann et al.'s [41] approach to random matrix modelling.

Drawing from the assumptions laid out in Section 6.1.1, the discussion proceeds to realisable models for Equations (6.8) to (6.10). Koch showed that tractable update formulas exist when the vector- and matrix-variate posterior distributions of Equation (6.4) are of the form

$$p(\mathbf{x}_k|\mathbf{X}_k, Z_{1:k}) = \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{k|k}, \mathbf{P}_{k|k} \otimes \mathbf{X}_k), \tag{6.12}$$

$$p(\mathbf{X}_k|Z_{1:k}) = \mathcal{IW}(\mathbf{X}_k; \nu_{k|k}, \mathbf{V}_{k|k}), \tag{6.13}$$

where $\otimes$ denotes the *Kronecker product*, $\mathbf{m}_{k|k}$ is the kinematic mean, $\mathbf{P}_{k|k}$ is the kinematic covariance in one spatial dimension, and $\mathcal{IW}(\mathbf{X}; \nu, \mathbf{V})$ denotes the *inverse Wishart* distribution [72] defined over the matrix $\mathbf{X}$ with degrees of freedom ν and SPD scale matrix $\mathbf{V}$. The inverse Wishart distribution is defined on positive definite matrices. Its degrees of freedom parameter provides a notion of the uncertainty in the scale matrix [72].

The Gaussian modelling of the target's kinematics implies Kalman filter-like operation. In fact, Koch's random matrix formulation additionally requires identical Gaussian dynamics in every spatial dimension. Essentially, the kinematic covariance is estimated in one dimension only, and the extent, by means of the Kronecker product, is used to 'shape' this covariance to be of full dimension. The same applies for the process noise covariance. Hence the mention of the covariance in 'one spatial dimension'. A complete understanding of this process is not required, but the discussion will refer back to it shortly.

A Bayesian solution requires that the structures of the densities in Equations (6.12) and (6.13) remain unchanged during the prediction and update steps. Koch proved that this is indeed possible for the respective densities if appropriate dynamic and measurement models are used. A Gaussian remains unchanged if propagated through a linear Gaussian transformation – a property exploited by the Kalman filter. Similarly, an inverse Wishart remains approximately unchanged when the transformation is a Wishart distribution. The prediction step maintains the required formats by using appropriate dynamic models, and the decoupled prediction procedure given by Equations (6.8) and (6.9).

The measurement model that results from the derivation is given by

$$h(Z_k|N_k, \mathbf{x}_k, \mathbf{X}_k) = \prod_{i=1}^{N_k} \mathcal{N}(\mathbf{z}_k^{(i)}; \mathbf{H}_k \mathbf{m}_{k|k-1}, \mathbf{X}_k) \tag{6.14}$$

$$\propto \mathcal{N}\left(\mathbf{z}_k; \mathbf{H}_k \mathbf{m}_{k|k-1}, \frac{\mathbf{X}_k}{N_k}\right) \mathcal{W}(\mathbf{Z}_k; N_k - 1, \mathbf{X}_k), \tag{6.15}$$

where $\mathbf{H}_k$ denotes the linear observation matrix, and $\mathcal{W}(\mathbf{Z}; \nu, \mathbf{V})$ denotes the *Wishart* distribution defined over the variable $\mathbf{Z}$ with degrees of freedom ν and scale matrix $\mathbf{V}$. The measurement model of Equation (6.15) constitutes the first factor of Equation (6.11). The second factor, $h(N_k|\mathbf{x}_k, \mathbf{X}_k)$, is approximated as a constant. The apparent factorisation again allows the required structure of the joint density to be maintained by means of a factored update step akin to the prediction [39]. Factoring of the kinematic and extent prediction and update equations does not invalidate joint kinematic-extent estimation, since the kinematic formulas remain conditioned on the target's extent.

As mentioned before, the form of the posterior in Equation (6.12) implies that the covariance of the kinematic distribution is shaped by the object's extent. The relation between the extent estimate and the covariance is the reason why the inverse Wishart distribution is the candidate distribution for the modelling of extent. Essentially, the algorithm infers about an unknown covariance using the extent's inverse Wishart distribution. The inverse Wishart distribution happens to be the conjugate prior for the covariance of a multi-variate normal distribution [73, p. 70], and is therefore suited for this particular application.

6.1.3 Incorporation of Statistical Sensor Noise

Two important shortcomings can be identified from the previous approach to extended target tracking. Firstly, the formulation does not allow full dimensional kinematic process noise covariance. More importantly, the measurement model of Equation (6.15) does not account for statistical sensor errors. Addressing these issues is the subject of the work of Feldmann et al. [41], which will now be presented.

The discussion begins with the prediction update proposed by Feldmann et al. The derivation relies on the assumption that the kinematic and extent estimates can be approximated as independent, thereby allowing the posterior of Equation (6.4) to be rewritten as

$$p(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k}) \approx p(\mathbf{x}_k|Z_{1:k}) p(\mathbf{X}_k|Z_{1:k}) = \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{k|k}, \mathbf{P}_{k|k}) \mathcal{IW}(\mathbf{X}_k; \nu_{k|k}, \mathbf{V}_{k|k}), \tag{6.16}$$

where $\mathbf{P}_{k|k}$ is now the full dimensional kinematic covariance. With the assumed independence, and further assuming independent dynamic models, standard Kalman filter prediction equations apply for the kinematics, i.e.

$$\mathbf{m}_{k|k-1} = \mathbf{F}_k \mathbf{m}_{k-1|k-1}, \tag{6.17}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k. \tag{6.18}$$

The above implies that the conditioning on the extent in Equation (6.8) may be removed, in other words $p(\mathbf{x}_k|\mathbf{X}_k, Z_{1:k-1}) = p(\mathbf{x}_k|Z_{1:k-1})$.

Assuming that the target extent does not tend to change over time, the recommended extent prediction is given by

$$\hat{\mathbf{X}}_{k|k-1} = \hat{\mathbf{X}}_{k-1|k-1}, \tag{6.19}$$

where $\hat{(\cdot)}$ denotes the expectation, i.e $\hat{\mathbf{X}}$ is the expected value of $\mathcal{IW}(\mathbf{X}; \nu, \mathbf{V})$. For an inverse Wishart distribution of the form $\mathcal{IW}(\mathbf{X}; \nu, \mathbf{V})$, the expected value is given by

$$\hat{\mathbf{X}} = \frac{\mathbf{V}}{\nu - d - 1}, \tag{6.20}$$

where d is the dimension of the square scale matrix $\mathbf{V}$. In terms of the distribution parameters, the prediction update is

$$\mathbf{V}_{k|k-1} = \frac{\alpha_{k|k-1}}{\alpha_{k-1|k-1}} \mathbf{V}_{k-1|k-1}, \tag{6.21}$$

where

$$\alpha_{(\cdot)} = \nu_{(\cdot)} - d - 1 \tag{6.22}$$

is a redundant variable comparable to the degrees of freedom ν, and is used only to simplify the recursive equations. For the discussion to follow ν will often be omitted, but the relationship in Equation (6.22) remains valid. An increased uncertainty in the extent would be expected following the prediction update. In view of this, the degrees of freedom parameter, which describes the precision of the corresponding expectation, is suggested to decrease according to

$$\alpha_{k|k-1} = 2 + \exp(-t_k/\tau)(\alpha_{k-1|k-1} - 2). \tag{6.23}$$

Equations (6.21) and (6.23) are derived from heuristic considerations. They are almost the exact replicas of the corresponding equations proposed by Koch [39], differing only slightly in the update of the degrees of freedom parameter. Important to realise is that, even though heuristic, the implied dynamic model retains the inverse Wishart form of the extent distribution. Koch also derived an alternative, more mathematically rigorous, prediction update in his original work, but stated that it does not offer much to gain in comparison with the heuristic approach [39].

Moving on to the measurement update, the likelihood function of Equation (6.14) can be rewritten to include sensor noise as

$$h(Z_k|N_k, \mathbf{x}_k, \mathbf{X}_k) \propto \mathcal{N}\left(\mathbf{z}_k; \mathbf{H}_k\mathbf{m}_{k|k-1}, \frac{\lambda\mathbf{X}_k + \mathbf{R}_k}{N_k}\right) \mathcal{W}(\mathbf{Z}_k; N_k - 1, \lambda\mathbf{X}_k + \mathbf{R}_k), \tag{6.24}$$

where $\mathbf{R}_k$ is the sensor noise covariance matrix, and λ is a scaling factor that serves as a means to spread the contribution of the object's extent [40]. Unfortunately, no appropriate conjugate prior exists for the likelihood of Equation (6.24) that is analytically tractable. To circumvent this issue, Feldmann et al. propose to approximate object extent as non-random during the measurement update, i.e. the uncertainty in $\mathbf{X}_k$ is ignored, and the term substituted with its expectation $\hat{\mathbf{X}}_k$. Consequently, the second factor of Equation (6.24) falls away, thereby enabling the use of standard Kalman filter equations in the kinematic measurement update. The resulting equations are given by

$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_{k|k-1}(\bar{Z}_k - \mathbf{H}_k\mathbf{m}_{k|k-1}), \tag{6.25}$$

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_{k|k-1}\mathbf{S}_{k|k-1}\mathbf{K}_{k|k-1}^T, \tag{6.26}$$

with

$$\mathbf{S}_{k|k-1} = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^T + \frac{\mathbf{Y}_{k|k-1}}{N_k}, \tag{6.27}$$

$$\mathbf{Y}_{k|k-1} = \lambda\hat{\mathbf{X}}_{k|k-1} + \mathbf{R}_k, \tag{6.28}$$
$$\mathbf{K}_{k|k-1} = \mathbf{P}_{k|k-1}\mathbf{H}_k^T\mathbf{S}_{k|k-1}. \tag{6.29}$$

The measurement update corrects for the extent estimate $\hat{\mathbf{X}}_k$, rather than the parameter $\mathbf{X}_k$, using an innovation and observation term:

$$\hat{\mathbf{X}}_{k|k} = \frac{1}{\alpha_{k|k}}(\alpha_{k|k-1}\hat{\mathbf{X}}_{k|k-1} + \breve{\mathbf{N}}_{k|k-1} + \breve{\mathbf{Z}}_{k|k-1}), \tag{6.30}$$

where

$$\breve{\mathbf{N}}_{k|k-1} = \hat{\mathbf{X}}_{k|k-1}^{1/2}\mathbf{S}_{k|k-1}^{-1/2}\mathbf{N}_{k|k-1}(\mathbf{S}_{k|k-1}^{-1/2})^T(\hat{\mathbf{X}}_{k|k-1}^{1/2})^T, \tag{6.31}$$
$$\breve{\mathbf{Z}}_{k|k-1} = \hat{\mathbf{X}}_{k|k-1}^{1/2}\mathbf{Y}_{k|k-1}^{-1/2}\tilde{Z}_k(\mathbf{Y}_{k|k-1}^{-1/2})^T(\hat{\mathbf{X}}_{k|k-1}^{1/2})^T. \tag{6.32}$$

are the symmetric and normalised counterparts of

$$\mathbf{N}_{k|k-1} = (\bar{Z}_k - \mathbf{H}_k\mathbf{m}_{k|k-1})(\bar{Z}_k - \mathbf{H}_k\mathbf{m}_{k|k-1})^T \tag{6.33}$$

and the measurement scattering matrix of Equation (6.3), representing the innovation and observation terms respectively. Equations (6.31) and (6.32) applies appropriate scaling to these respective terms, and ensures that the matrices remain symmetric positive definite. The degrees of freedom parameter is updated according to

$$\alpha_{k|k} = \alpha_{k|k-1} + N_k, \tag{6.34}$$

signifying increased precision in the extent estimate.

The update equations laid out in this section maintain the structure of the posterior in Equation (6.16). The reader may be poised to remark that the algorithm of Feldmann et al. does away with joint kinematic-extent estimation, due to the required independence assumptions. Although this may be true in a strict Bayesian sense, the required interdependency between the kinematic and extent estimation is provided by the mean innovation term $\mathbf{N}_{k|k-1}$ in Equation (6.33), and the innovation covariance $\mathbf{S}_{k|k-1}$ of Equation (6.27) [41]. Moreover, the technique proved to be more accurate in the presence of marked sensor noise [41]. The simplicity of the eventual update equations is also an important consideration for real-time applications.

6.2 Gaussian Inverse Wishart Probability Hypothesis Density Filter

The preceding discussions culminate in a tracking formulation proposed by the author that considers multiple extended targets. The algorithm combines the framework of the GM-PHD filter with the random matrix approach of Feldmann et al. The resulting estimator, termed the *Gaussian inverse Wishart probability hypothesis density* (GIW-PHD) filter, allows the tracking of multiple extended targets without the need for explicit measurement-to-track assignments.

Considering a procedure alike to the presentation of previous recursive estimators, recall the generic form for the propagation of lower order statistical moments of the multi-target state RFS given by

$$\upsilon(\mathbf{x}_k|Z_{1:k-1}) = \int p_S(\mathbf{x}_{k-1}) f(\mathbf{x}_k|\mathbf{x}_{k-1}) \upsilon(\mathbf{x}_{k-1}|Z_{1:k-1}) d\mathbf{x}_{k-1} + \int \beta(\mathbf{x}_k|\mathbf{x}_{k-1}) \upsilon(\mathbf{x}_{k-1}|Z_{1:k-1}) d\mathbf{x}_{k-1} + \gamma(\mathbf{x}_k), \quad \text{((5.5) revisited)}$$

$$\upsilon(\mathbf{x}_k|Z_{1:k}) = [1 - p_D(\mathbf{x}_k)]\upsilon(\mathbf{x}_k|Z_{1:k-1}) + \sum_{\mathbf{z}\in Z_k} \frac{p_D(\mathbf{x}_k) h(\mathbf{z}|\mathbf{x}_k) \upsilon(\mathbf{x}_k|Z_{1:k-1})}{\kappa(\mathbf{z}) + \int p_D(\mathbf{x}_k) h(\mathbf{z}|\mathbf{x}_k) \upsilon(\mathbf{x}_k|Z_{1:k-1}) d\mathbf{x}_k}. \quad \text{((5.6) revisited)}$$

The goal is to incorporate the random matrix modelling techniques of Section 6.1.3 within the PHD framework of Equations (5.5) and (5.6). Before presenting the closed-form recursive equations, the GM-PHD filter assumptions of Section 5.2 need to be revisited. The modelling techniques described in the previous section suggest the following amendments.

A.1: The temporal kinematic and extent evolution of each target is described by decoupled dynamic models, i.e.

$$f(\mathbf{x}_k, \mathbf{X}_k|\mathbf{x}_{k-1}, \mathbf{X}_{k-1}) = f(\mathbf{x}_k|\mathbf{x}_{k-1}) f(\mathbf{X}_k|\mathbf{X}_{k-1}). \tag{6.35}$$

A.2: The kinematic evolution is a linear Gaussian transformation, and the extent is required to remain inverse Wishart distributed, i.e.

$$f(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k; \mathbf{F}_k \mathbf{m}_{k-1|k-1}, \mathbf{Q}_k), \tag{6.36}$$
$$f(\mathbf{X}_k|\mathbf{X}_{k-1}) : \mathcal{IW}(\mathbf{X}_{k-1}; \nu_{k-1|k-1}, \mathbf{V}_{k-1|k-1}) \to \mathcal{IW}(\mathbf{X}_k; \nu_{k|k-1}, \mathbf{V}_{k|k-1}). \tag{6.37}$$

A.3: The extent is deterministic during the measurement update, and the measurement model is described by a linear Gaussian distribution as follows

$$h(\bar{Z}_k|\mathbf{x}_k, \mathbf{X}_k) = \mathcal{N}\left(\mathbf{z}_k; \mathbf{H}_k \mathbf{m}_{k|k-1}, \frac{\lambda \hat{\mathbf{X}}_{k|k-1} + \mathbf{R}_k}{N_k}\right). \tag{6.38}$$

A.4: The intensities of the spawn and spontaneous birth RFSs are decoupled Gaussian inverse Wishart (GIW) mixtures of the form

$$\beta(\mathbf{x}_k, \mathbf{X}_k|\mathbf{x}_{k-1}, \mathbf{X}_{k-1}) = \sum_{i=1}^{J_{\beta,k}} w_{\beta,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{F}_{\beta,k}^{(i)} \mathbf{m}_{k-1|k-1} + \mathbf{d}_{\beta,k}^{(i)}, \mathbf{Q}_{\beta,k}^{(i)}) \mathcal{IW}(\mathbf{X}_k; \nu_{\beta,k}^{(i)}, \mathbf{V}_{\beta,k}^{(i)}), \tag{6.39}$$

$$\gamma(\mathbf{x}_k, \mathbf{X}_k) = \sum_{i=1}^{J_{\gamma,k}} w_{\gamma,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{\gamma,k}^{(i)}, \mathbf{P}_{\gamma,k}^{(i)}) \mathcal{IW}(\mathbf{X}_k; \nu_{\gamma,k}^{(i)}, \mathbf{V}_{\gamma,k}^{(i)}). \tag{6.40}$$

The discussion proceeds to the implementation details with the aforementioned assumptions in mind. Guided by the Gaussian inverse Wishart representation for the posterior of Equation (6.16), assume that the posterior intensity at time $k-1$ is a GIW mixture of the form

$$\upsilon(\mathbf{x}_{k-1}, \mathbf{X}_{k-1}|Z_{1:k-1}) = \sum_{i=1}^{J_{k-1|k-1}} w_{k-1|k-1}^{(i)} \mathcal{N}(\mathbf{x}_{k-1}; \mathbf{m}_{k-1|k-1}^{(i)}, \mathbf{P}_{k-1|k-1}^{(i)})$$

$$\mathcal{IW}(\mathbf{X}_{k-1}; \nu^{(i)}_{k-1|k-1}, \mathbf{V}^{(i)}_{k-1|k-1}). \quad (6.41)$$

Following the presentation format of Section 5.2, the intensity that succeeds the prediction update is of the form

$$\upsilon(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) = \upsilon_S(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) + \upsilon_\beta(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) + \upsilon_\gamma(\mathbf{x}_k, \mathbf{X}_k), \quad (6.42)$$

where the GIW mixture

$$\upsilon_S(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) = p_S \sum_{i=1}^{J_{k-1|k-1}} w^{(i)}_{k-1|k-1} \mathcal{N}(\mathbf{x}_k; \mathbf{m}^{(i)}_{S,k|k-1}, \mathbf{P}^{(i)}_{S,k|k-1}) \mathcal{IW}(\mathbf{X}_k; \nu^{(i)}_{S,k|k-1}, \mathbf{V}^{(i)}_{S,k|k-1}) \quad (6.43)$$

with

$$\mathbf{V}^{(i)}_{S,k|k-1} = \frac{\alpha^{(i)}_{S,k|k-1}}{\alpha^{(i)}_{k-1|k-1}} \mathbf{V}^{(i)}_{k-1|k-1}, \quad (6.44)$$

$$\alpha^{(i)}_{S,k|k-1} = 2 + \exp(-t_k/\tau)(\alpha^{(i)}_{k-1|k-1} - 2), \quad (6.45)$$

models surviving targets. The kinematic update equations of υ_S are the same as those given in Section 5.2. The mixture

$$\upsilon_\beta(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) = \sum_{i=1}^{J_{k-1|k-1}} \sum_{j=1}^{J_{\beta,k}} w^{(i)}_{k-1|k-1} w^{(j)}_{\beta,k} \mathcal{N}(\mathbf{x}_k; \mathbf{m}^{(i,j)}_{\beta,k|k-1}, \mathbf{P}^{(i,j)}_{\beta,k|k-1}) \mathcal{IW}(\mathbf{X}_k; \nu^{(i)}_{\beta,k|k-1}, \mathbf{V}^{(i)}_{\beta,k|k-1}) \quad (6.46)$$

models targets that spawn from existing ones. Again, the kinematic update equations of υ_β are the same as those given in Section 5.2. The extent parameters $\nu^{(i)}_{\beta,k|k-1}$ and $\mathbf{V}^{(i)}_{\beta,k|k-1}$ may be chosen to represent general elliptical shapes, or may alternatively include information from the parent mixture.

The spontaneous birth intensity $\upsilon_\gamma(\mathbf{x}_k, \mathbf{X}_k)$ is given by Equation (6.40). Upon combining the respective intensities, the predicted intensity of Equation (6.42) remains a GIW mixture and is of the form

$$\upsilon(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) = \sum_{i=1}^{J_{k|k-1}} w^{(i)}_{k|k-1} \mathcal{N}(\mathbf{x}_k; \mathbf{m}^{(i)}_{k|k-1}, \mathbf{P}^{(i)}_{k|k-1}) \mathcal{IW}(\mathbf{X}_k; \nu^{(i)}_{k|k-1}, \mathbf{V}^{(i)}_{k|k-1}). \quad (6.47)$$

Moving on to the measurement update, Mahler [74] suggests that the use of cluster measurements dictate a slightly altered formulation compared to Equation (5.22). The cluster measurement update and resulting posterior is given by

$$\upsilon(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k}) = (1 - p_D)\upsilon(\mathbf{x}_k, \mathbf{X}_k|Z_{1:k-1}) + \sum_{\langle Z_k \rangle} \sum_{W \in \langle Z_k \rangle} \upsilon_D(\mathbf{x}_k, \mathbf{X}_k; W), \quad (6.48)$$

where Z_K denotes the finite set of measurements at time k, and the notation $\langle Z \rangle$ denotes a partition[1] of the set Z. The use of the partitioning operator in conjunction with a summation sign indicates that the summation is over all possible partitions. The elements

[1] A partition of set S is a collection of disjoint subsets of S whose union is S [75].

of any particular partition are themselves sets and in this case they resemble cluster measurements. Measurement partitioning methods will be discussed in a subsequent chapter. The detection mixture υ_D is given by

$$\upsilon_D(\mathbf{x}_k, \mathbf{X}_k; W) = \sum_{i=1}^{J_{k|k-1}} w_{k|k}^{(i)}(W)\mathcal{N}(\mathbf{x}_k; \mathbf{m}_{k|k}^{(i)}(W), \mathbf{P}_{k|k}^{(i)}(W))\mathcal{IW}(\mathbf{X}_k; \nu_{k|k}^{(i)}(W), \mathbf{V}_{k|k}^{(i)}(W)), \tag{6.49}$$

with

$$w_{k|k}^{(i)}(W) = \frac{p_D w_{k|k-1}^{(i)} q_k^{(i)}(W)}{\kappa_k(W) + p_D \sum_{j=1}^{J_{k|k-1}} w_{k|k-1}^{(j)} q_k^{(j)}(W)}, \tag{6.50}$$

$$\mathbf{m}_{k|k}^{(i)}(W) = \mathbf{m}_{k|k-1}^{(i)} + \mathbf{K}_{k|k-1}^{(i)}(\bar{W} - \mathbf{H}_k \mathbf{m}_{k|k-1}^{(i)}), \tag{6.51}$$

$$\mathbf{P}_{k|k}^{(i)}(W) = \mathbf{P}_{k|k-1}^{(i)} - \mathbf{K}_{k|k-1}^{(i)} \mathbf{S}_{k|k-1}^{(i)} (\mathbf{K}_{k|k-1}^{(i)})^T, \tag{6.52}$$

$$\mathbf{K}_{k|k-1}^{(i)} = \mathbf{P}_{k|k-1}^{(i)} \mathbf{H}_k (\mathbf{S}_{k|k-1}^{(i)})^{-1}, \tag{6.53}$$

$$\mathbf{S}_{k|k-1}^{(i)} = \mathbf{H}_k \mathbf{P}_{k|k-1}^{(i)} \mathbf{H}_k^T + \frac{\mathbf{Y}_{k|k-1}^{(i)}}{N_W}, \tag{6.54}$$

$$\mathbf{Y}_{k|k-1}^{(i)} = \lambda \hat{\mathbf{X}}_{k|k-1}^{(i)} + \mathbf{R}_k, \tag{6.55}$$

$$\alpha_{k|k}^{(i)}(W) = \alpha_{k|k}^{(i)} + N_W, \tag{6.56}$$

$$\mathbf{V}_{k|k}^{(i)}(W) = \frac{1}{\alpha_{k|k}^{(i)}}(\alpha_{k|k-1}^{(i)} \hat{\mathbf{X}}_{k|k-1}^{(i)} + \breve{\mathbf{N}}_{k|k-1}^{(i)} + \breve{\mathbf{W}}_{k|k-1}^{(i)}), \tag{6.57}$$

$$\breve{\mathbf{N}}_{k|k-1}^{(i)} = (\hat{\mathbf{X}}_{k|k-1}^{(i)})^{1/2} (\mathbf{S}_{k|k-1}^{(i)})^{-1/2} \mathbf{N}_{k|k-1}^{(i)} ((\mathbf{S}_{k|k-1}^{(i)})^{-1/2})^T ((\hat{\mathbf{X}}_{k|k-1}^{(i)})^{1/2})^T, \tag{6.58}$$

$$\breve{\mathbf{W}}_{k|k-1} = (\hat{\mathbf{X}}_{k|k-1}^{(i)})^{1/2} (\mathbf{Y}_{k|k-1}^{(i)})^{-1/2} \tilde{W} ((\mathbf{Y}_{k|k-1}^{(i)})^{-1/2})^T ((\hat{\mathbf{X}}_{k|k-1}^{(i)})^{1/2})^T, \tag{6.59}$$

$$\mathbf{N}_{k|k-1}^{(i)} = (\bar{W} - \mathbf{H}_k \mathbf{m}_{k|k-1})(\bar{W} - \mathbf{H}_k \mathbf{m}_{k|k-1})^T, \tag{6.60}$$

where N_W is the number of measurements in the partition W. The first and second terms in Equation (6.48) again account for missed detections and mixture components resulting from new measurements, respectively. For the kinematic part, the recursive update is fundamentally the same as in the standard Kalman filter, apart from the innovation covariance $\mathbf{S}_{k|k-1}^{(i)}$. The extent is updated in the same manner as laid out in Section 6.1.3. The factor $q_k^{(i)}(W)$ may be calculated as before using the measurement likelihood. For the random matrix extent model, it is suggested to use

$$q_k^{(i)}(W) = \mathcal{N}(\bar{W}; \mathbf{H}_k \mathbf{m}_{k|k-1}^{(i)}, \mathbf{S}_{k|k-1}^{(i)}) \tag{6.61}$$

as the measurement likelihood factor. In keeping with the assumption that the extent is non-random during the measurement update, Equation (6.61) does not account for the uncertainty in the extent distribution.

Equations (6.42) and (6.48) are the high-level representation of the GIW-PHD filter's recursive equations. The filter explicitly accounts for extended targets, while allowing for a Gaussian representation of uncertainty for both target kinematic states and sensor error. Furthermore, multiple extended targets can be tracked simultaneously without explicit data association. The added benefit of target extent information may be of

great importance to higher level decision making entities. The proposed GIW-PHD filter deviates from the one presented by Granström and Orguner [76] in the same way the separate random matrix formulations [39, 41] differ with regards to uncoupling. To the author's knowledge, such an embedding of Feldmann's [41] random matrix model in a GM-PHD filtering framework has not been implemented before.

Chapter 7

Data Fusion Architecture

The theoretical development of Chapters 5 and 6 culminates in a description of their use in the proposed radar-vision data fusion implementation. Up until now, the focus was on the measurement extraction and tracking facets of DATMO. This chapter will discuss the way in which data fusion aids the detection and tracking process. The entire end-to-end data flow will also be presented. An overview of how the various entities comprising the overall system fit together will firstly be given. The discussion will then progress to the details of the data fusion algorithm, and how the information is used to bring about the eventual system output.

7.1 Algorithm Overview

Before delving into detail about the parts that constitute the algorithm, a discussion of the basic workflow should prove helpful. The data fusion framework is illustrated by the flow diagram in Figure 7.1. The processing chain on the right shows the various steps that are needed to extract cluster measurements from the stereo vision subsystem, namely feature detection, feature tracking and DBSCAN clustering. The details with regard to vision-based measurement extraction were given in Section 4.1.

As shown in Figure 7.1, radar measurements are not passed directly to the data fusion block. This is consistent with most radar-vision fusion approaches found in literature. Instead, radar-based state estimation precede the fusion step. The reason for doing so is twofold. Due to unsynchronised reporting from the respective sensors, raw radar and camera measurements cannot simply be combined at each time step. In addition, raw measurements from the radar contain insufficient Doppler information for identifying moving objects. In highway driving conditions, and even in pedestrian environments, surrounding objects' relative radial velocity will often exceed a mere $\pm 1.7\,\mathrm{m/s}$ unambiguous interval. State estimation addresses these problems: Filtering allows the prediction of state estimates to arbitrary time instances, while the inferred velocities can be used to filter stationary tracks. Section 7.2 will describe radar target tracking in more detail.

The fusion block receives the state estimates from the radar's GM-PHD filter, along with the cluster measurements extracted by the vision algorithm. The process of combining this data is the subject of Section 7.3. Data fusion results in a fused measurement set that is subsequently passed to the GIW-PHD extended target tracker. The state estimates of the GIW-PHD filter resemble the eventual output of the DATMO system.

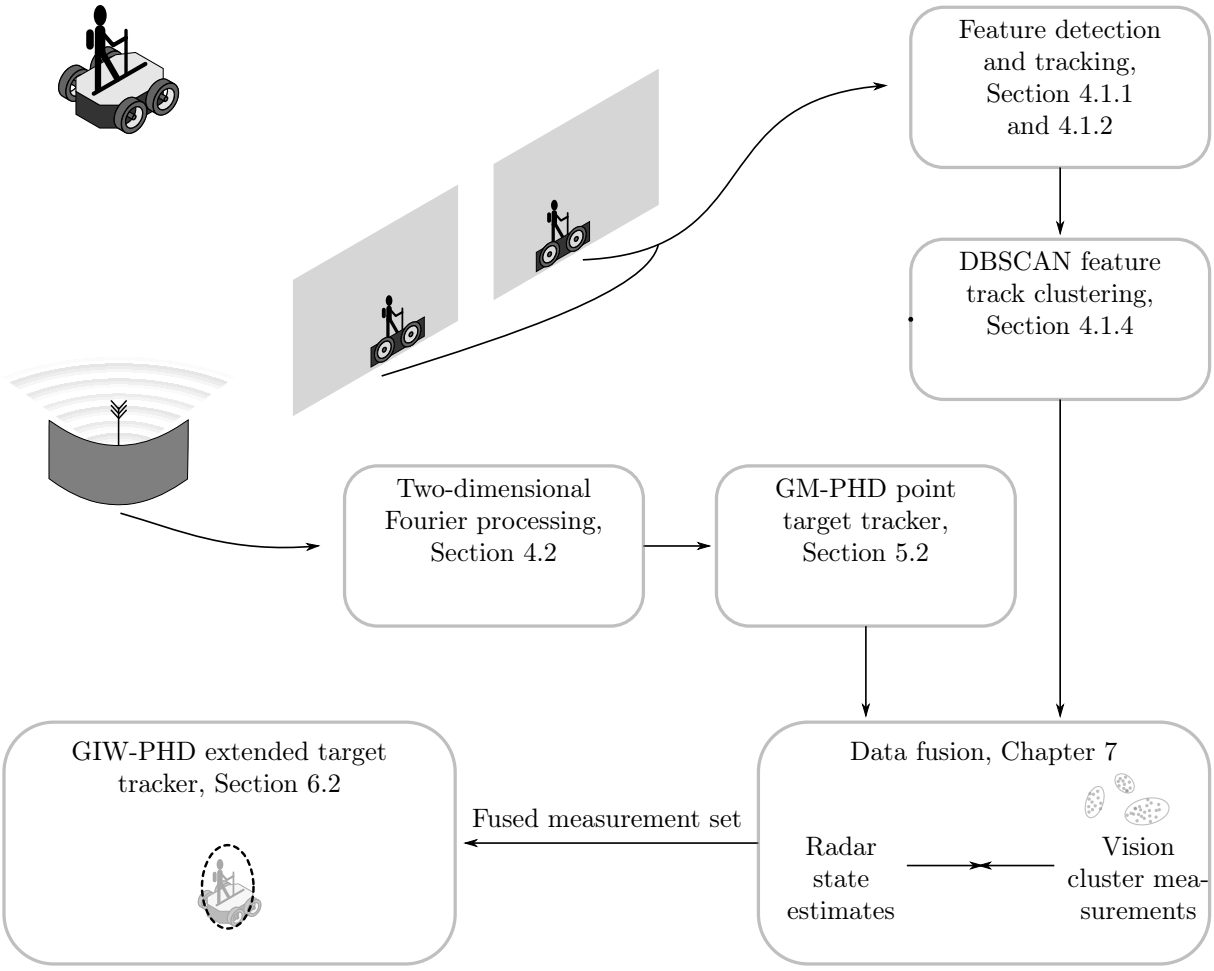


Figure 7.1: Flow diagram of the proposed DATMO system.

7.2 Radar Target Tracking

For reasons mentioned in the previous section, radar measurements are tracked before being passed to the fusion block. The environments under consideration require the tracking of multiple targets. Moreover, the limited resolution capabilities of the relevant hardware prescribes the use of the standard point object assumption for radar target tracking. For these reasons, the GM-PHD filter described in Section 5.2 is chosen for radar-based state estimation. The closed-form GM-PHD formulas laid out in Section 5.2 are therefore directly applicable for the radar-based recursive state estimation. The following sections will detail the various models implemented to realise the filter.

7.2.1 Pruning and Merging

A key step of both PHD filter variants that have been omitted up until now is pruning and merging. Pruning and merging is not so much related to the theoretical functionality of the PHD filter, but it is indispensable for limiting the computation requirements of the filter. Consider the high-level formulation of the GM-PHD recursion given by Equations (5.14) and (5.22), which are repeated here for convenience

$$v(\mathbf{x}_k|Z_{1:k-1}) = v_S(\mathbf{x}_k|Z_{1:k-1}) + v_\beta(\mathbf{x}_k|Z_{1:k-1}) + v_\gamma(\mathbf{x}_k), \qquad \text{((5.14) revisited)}$$

Algorithm 1 GM-PHD pruning and merging. Adapted from Vo and Ma [33] to include identity tracking.

Require: The identity-augmented state set $X_k = \{w_k^{(i)}, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)}, \mathrm{id}_k^{(i)}\}_{i=1}^{J_k}$, the identity set at the previous time step $A_{k-1} = \{\mathrm{id}_{k-1}^{(i)}\}_i^{J_{k-1}}$, a truncation threshold T, and a merging threshold U.

$l \leftarrow 0$, $I \leftarrow \{i = 1, \ldots, J_k | w_k^{(i)} > T\}$
while $I \neq \emptyset$ **do**
  $l \leftarrow l + 1$
  $j \leftarrow \arg\max_{i \in I} w_k^{(i)}$
  $L \leftarrow \{i \in I | \mathcal{B}(\mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)}, \mathbf{m}_k^{(j)}, \mathbf{P}_k^{(j)}) \leq U\}$ ▷ Indices to merge
  ▷ Calculate merged weight, mean and covariance
  $\bar{w}_k^{(l)} \leftarrow \sum_{i \in L} w_k^{(i)}$
  $\bar{\mathbf{m}}_k^{(l)} \leftarrow \frac{1}{\bar{w}_k^{(l)}} \sum_{i \in L} w_k^{(i)} \mathbf{m}_k^{(i)}$
  $\bar{\mathbf{P}}_k^{(l)} \leftarrow \frac{1}{\bar{w}_k^{(l)}} \sum_{i \in L} w_k^{(i)} (\mathbf{P}_k^{(i)} + (\bar{\mathbf{m}}_k^{(l)} - \mathbf{m}_k^{(i)})(\bar{\mathbf{m}}_k^{(l)} - \mathbf{m}_k^{(i)})^T)$
  ▷ Determine index
  $b \leftarrow \arg\max_{i \in L}(w_k^{(i)})$
  $B \leftarrow \{i \in L | \mathrm{id}_k^{(i)} \in A_{k-1}\}$
  if $B \neq \emptyset$ **then**
    $b \leftarrow \arg\max_{i \in B}(w_k^{(i)})$
  end if
  $\bar{\mathrm{id}}_k^{(l)} \leftarrow \mathrm{id}_k^{(b)}$
  $I \leftarrow I \backslash L$
end while
return The pruned and merged identity-augmented Gaussian mixture components $\{\bar{w}_k^{(i)}, \bar{\mathbf{m}}_k^{(i)}, \bar{\mathbf{P}}_k^{(i)}, \bar{\mathrm{id}}_k^{(l)}\}_{i=1}^{l}$.

$$v(\mathbf{x}_k | Z_{1:k}) = (1 - p_D) v(\mathbf{x}_k | Z_{1:k-1}) + \sum_{\mathbf{z} \in Z_k} v_D(\mathbf{x}_k; \mathbf{z}). \qquad \text{((5.22) revisited)}$$

Important to note is that, following each update step, the posterior intensity $v(\mathbf{x}_k | Z_{1:k})$ contains $(J_{k-1|k-1}(1 + J_{\beta,k}) + J_{\gamma,k})(1 + |Z_k|)$ mixture components, where $|Z_k|$ is the number of measurements in the measurement set at time k. To limit the boundless increase of components, a pruning and merging strategy must be implemented. Pruning is the deletion of target tracks, and entails the removal of mixture components from the target state set. The merging procedure combines similar tracks into a single track.

A slight deviation from the method presented by Vo and Ma [33] is used. Firstly, no constraint is placed on the maximum allowable number of mixture components, and the Bhattacharyya distance[1] $\mathcal{B}(\cdot, \cdot, \cdot, \cdot)$ is used as merging criteria instead of the Mahalanobis distance, since it is a symmetric measure [77]. The Bhattacharyya distance between

[1]The Bhattacharyya distance between two Gaussian distributions $\mathcal{N}(\mathbf{x}_1; \mathbf{m}_1, \mathbf{P}_1)$ and $\mathcal{N}(\mathbf{x}_2; \mathbf{m}_2, \mathbf{P}_2)$ is given by the closed-form expression $\mathcal{B}(\mathbf{m}_1, \mathbf{P}_1, \mathbf{m}_2, \mathbf{P}_2) = \frac{1}{8}\mathbf{u}^T \boldsymbol{\Gamma}^{-1} \mathbf{u} + \frac{1}{2} \ln(|\mathbf{P}_1|^{-0.5} |\mathbf{P}_2|^{-0.5} |\boldsymbol{\Gamma}|)$, where $\mathbf{u} = \mathbf{m}_1 - \mathbf{m}_2$, and $\boldsymbol{\Gamma} = \frac{1}{2}\mathbf{P}_1 + \frac{1}{2}\mathbf{P}_2$ [77].

two Gaussian distributions considers the uncertainty of both distributions, whereas the Mahalanobis distance does not. It is therefore favoured for the calculation of probabilistic distance (dissimilarity) measures.

The pseudo code for pruning and merging with identity tracking is shown in Algorithm 1. The process is initiated by pruning, which simply involves truncating all mixtures with weights below a certain threshold. Thereafter, mixture components are merged by comparing the Bhattacharyya distance to a merging threshold. The equations for reducing numerous Gaussian mixture components into a single component are given by lines 6 to 8 in Algorithm 1. When merging, an attempt is made to find mixtures in the current merge set that existed at the previous time step. If found, the index of the component with the highest weight among them is used as the index for the merged component. This simple procedure enables the target identity to be tracked. The process of merging mixture components is also referred to as mixture reduction.

7.2.2 Gaussian Mixture Models

In order to close the discussion on the radar multi-target tracker, the mixture models implemented for target births, spawning, surviving targets, and detected targets must be specified. With reference to the states and matrices in Equations (5.15) to (5.17) describing the dynamic model of surviving targets, the following applies: each target's motion is modelled by linear Gaussian dynamics in inertial space according to the constant velocity model [63]

$$\mathbf{x} = [x, y, z, v_x, v_y, v_z]^T, \tag{7.1}$$

$$\mathbf{F} = \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix} \otimes \mathbf{I}_3, \quad \mathbf{Q} = \sigma_w^2 \begin{bmatrix} \frac{1}{4}\Delta T^4 & \frac{1}{2}\Delta T^3 \\ \frac{1}{2}\Delta T^3 & \Delta T^2 \end{bmatrix} \otimes \mathbf{I}_3, \tag{7.2}$$

where σ_w is the acceleration noise standard deviation, ΔT is the time step, and $\mathbf{I}_d$ is the $d \times d$ identity matrix. The constant velocity model is sufficient for motion modelling in DATMO applications, since targets do not tend to deviate significantly from straight and consistent motion. Moreover, the resulting state space dimensionality is only twice the number of tracking axes, enabling efficient filters to be realised.

Measurements from the radar are in two-dimensional polar coordinates, therefore requiring non-linear update equations. The measurement model maps the target states in world reference frame Cartesian coordinates to noisy measurements in radar reference frame polar coordinates, i.e.,

$$h : \mathbf{x}^W \rightarrow \mathcal{N}(\mathbf{z}^R; \bar{\mathbf{z}}, \mathbf{R}), \tag{7.3}$$

with

$$\mathbf{z}^R = [r, \alpha]^T, \tag{7.4}$$

$$\mathbf{R} = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\alpha^2 \end{bmatrix}, \tag{7.5}$$

where $\mathbf{R}$ is the range-bearing noise covariance matrix, and σ_r and σ_α are the standard deviations in range and angle respectively. The elements of the measurement set Z_k used during the update step (see Equation (5.22)) are in the range-bearing format described in Equation (7.4). The unscented transform [62] is used to approximate non-linear transformations of Gaussian mixture components. Incorporation of the non-linear measurement

model results in a multi-target filter that is analogous to the unscented Kalman filter (UKF). To use the unscented transform in the GM-PHD update step, Equations (5.25) to (5.28) of the detection mixture is adapted according to UKF measurement update principles. The interested reader is referred to the work of Vo and Ma [33] for the exact update equations.

The only additional information required to close the discussion on the radar estimator is with regard to the format of the spawn and birth Gaussian mixtures. The chosen implementation disregards spontaneous target births, and reformulates the target spawning mixture $\upsilon_\beta(\mathbf{x}_k|Z_{1:k-1})$ to use measurements rather than targets to spawn mixture components. The target spawning mixture is constructed from a set that contains "unused" measurements of the previous time step, i.e.

$$Z_{\beta,k-1} = \{\mathbf{z}_{k-1}^{(i)}\}_{i=1}^{N_{\beta,k-1}}. \tag{7.6}$$

What is meant by "unused" will be clarified shortly. The spawn mixture is subsequently given by

$$\upsilon_\beta(\mathbf{x}_k|Z_{\beta,k-1}) = \sum_{i=1}^{N_{\beta,k-1}} w_{\beta,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{\beta,k}^{(i)}, \mathbf{P}_{\beta,k}^{(i)}), \tag{7.7}$$

with

$$\mathbf{m}_{\beta,k}^{(i)} = h^{-1}\left(Z_{\beta,k-1}^{(i)}\right), \tag{7.8}$$

$$h^{-1} : \mathbf{z}^R \to \mathbf{x}^W, \tag{7.9}$$

where $h^{-1}(\cdot)$ maps a range-bearing measurement in the RRF to the WRF state space. The covariance matrix assigned to newly spawned components is set according to a heuristic consideration. More specifically, it is chosen to be fairly large so as to cover a reasonably sized region where new targets are expected to emerge from.

During the update step of the GM-PHD recursion, a single measurement will spawn a new mixture component for every component present in the prior intensity. The weight of this mixture is calculated by evaluating the component's innovation density at the measurement position (see Equation (5.29)). When the observation is far from the prior mixture component in probabilistic terms, the resulting weight may be lower than the truncation threshold, and the newly spawned component will be removed immediately upon pruning. If this condition holds true for all components in the prior intensity, we define the measurement as "unused" and add it to $Z_{\beta,k}$. Stated otherwise, a measurement that has no effect on the merged posterior distribution is added to $Z_{\beta,k}$.

This paragraph explains the rationale behind the proposed reformulation of the spawn mixture. In the PHD filter evolution model of Section 5.1.1, new targets can only originate in two ways: either from existing mixture components or from the birth model. The birth and spawn mixture examples given by the GM-PHD authors are somewhat prohibitive, since the birth mixtures are hard-coded to cover parts of the surveillance area and the spawn mixture requires pre-existing components to produce new ones. For multi-target tracking in known and structured environments, simple birth mixture strategies will most definitely suffice. However, when the sensing platform moves through an ever changing environment, the process of formulating comprehensive birth mixtures becomes complex. Simply covering the surveillance area with many mixture components with large uncertainties is not recommended, since they will interact with the components that represent

actual targets. Unwanted interaction may reduce the estimation accuracy and induce an increased probability of identity mismatch errors. Furthermore, no more mixture components should be spawned than that is necessary, since the complexity of the recursion is highly correlated to the number of mixture components. The adaptive spawn mixture proposed earlier provides a versatile solution to target spawning. It inherently eludes the spawning of new components near existing targets and it is quick to react to areas where new targets emerge.

In the GM-PHD filtering framework, target state information is directly available from the target state set X_k. The state estimates that are considered for subsequent use may be limited to include only mixture components whose weight exceeds a certain threshold in order not to waste resources on weak tracks. In addition to the target states, PHD filters also provide an estimate of the expected number of targets. The number of targets, or cardinality, is calculated by summing the weights in the state set.

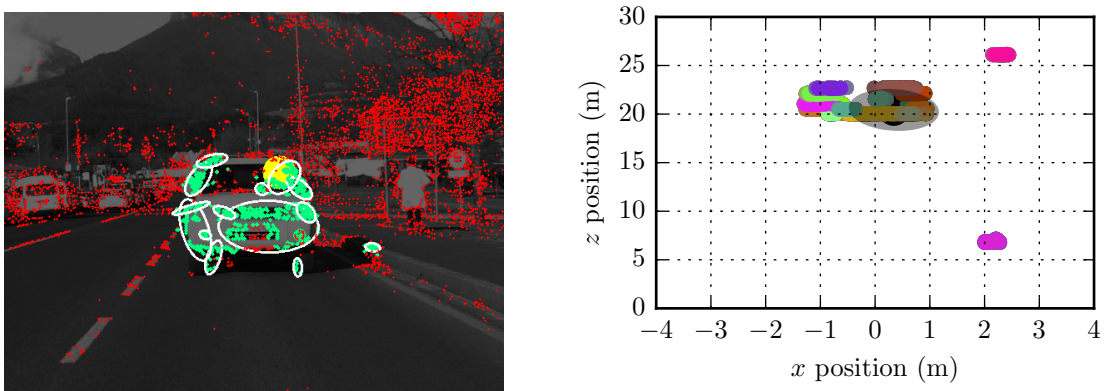
7.3 Track-to-Track Fusion

This section describes the way in which information from the vision and radar subsystems are combined. The proposed method is motivated by the characteristics of the measurement clusters output by the feature tracking framework. Image feature clustering deliberately aims at over-segmenting an object by choosing a small distance threshold for the DBSCAN algorithm. This is done in order to eliminate the majority of outliers that are not removed in the feature tracker. To clarify, point tracks that lie in a non-dense space are considered to be outliers, which indeed is often the case. Figure 7.2a shows an example of an over-segmented vehicle. It is clear that the moving object is successfully detected, but the information portrayed by individual cluster measurements covering it are not of much use in isolation. Furthermore, disjoint measurement clusters are not suited for the random matrix extended target observation model [76]. A subsequent grouping method is therefore required to indicate which clusters belong to the same object. However, accurately inferring group structures from the available data may be very difficult in the event that they are separated by large textureless regions. The vision measurement extraction algorithm is bound to suffer in this regard due to the absence of smoothing constraints in the scene flow calculations (contrary to dense scene flow algorithms).

With the aforementioned in mind, it is chosen not to fuse the radar and vision measurements by means of separate update formulations in a centralised Bayesian state estimator as described in Section 2.5.1. The supporting argument is that, for the proposed algorithm, more is gained if the information of both sensors is used to group over-segmented points belonging to a single object. These groups can then be tracked more reliably to provide the eventual estimates of moving objects.

7.3.1 Data Pre-Processing

To begin the presentation of the fusion algorithm, recall the information that is available at the fusion centre at each time step, namely the state estimates from the radar's GM-PHD filter and the cluster measurements extracted by the vision algorithm. The notation $X_{\mathrm{rdr},k}$ and $Z_{\mathrm{cam},k}$ is introduced to refer to the radar state set and vision cluster measurement set at time k, respectively. The data fusion algorithm attempts to exploit the information from the two subsystems in order to achieve superior overall perception performance. The



(a) (b)

Figure 7.2: An illustration of the over-segmented clusters extracted by the vision detection algorithm. (a) The groups of green points represent the cluster measurements, while red points are either DBSCAN outliers or stationary clusters. The yellow circle coincides with the mean of a radar mixture component. (b) The 2-D CRF projection of the same data. Cluster measurements are drawn with a random colour and the radar mixture is shown by the grey ellipse.

fact that both $X_{\text{rdr},k}$ and $Z_{\text{cam},k}$ result from some form of state estimation means that the algorithm may be classified as a track-to-track fusion variant (see Section 2.5.3).

As mentioned in the discussion above, the fusion algorithm should attempt to group over-segmented cluster measurements. This is done by associating cluster measurements to radar tracks to form grouped clusters. In classical track-to-track fusion methods, the densities describing target states are available for all the sensors involved. The feature tracking framework, however, tracks in the image coordinates. Hence, the uncertainty in the world reference frame (WRF) is not directly available for comparison with the radar's mixture components. Instead of proceeding with a deterministic association procedure, approximated Gaussian distributions are assigned to the cluster measurements using a stereo vision error model.

The process is initiated by projecting the radar mixture components and vision clusters from the WRF to the camera reference frame (CRF). Note that only radar mixture components with a mean velocity exceeding a certain threshold value is considered during fusion. The CRF is considered for the following reasons: Firstly, the upright dimension is disregarded during track fusion because no height information is available from the radar. Neglecting height is not expected to adversely effect the performance of fusion, since moving objects in DATMO environments seldom coexist above or below each other. Knowledge of the upright direction, however, is dependent on the current robot pose, which in turn is expected to drift. Consequently, it is required to perform the projection to 2-D in the CRF. Projecting to the CRF also facilitates the construction of approximate Gaussian densities that represent cluster measurements.

The stereo error approximation method described in an application note [78] of the camera manufacturer is used to fit Gaussian distributions to the vision cluster measurements. The error in depth is given by

$$\partial z = \frac{z^2}{fb}\partial d, \tag{7.10}$$

where z is the z position in the CRF, f is the camera focal length, b is the stereo baseline, and ∂d is the error in disparity. It is argued that the extent of a cluster measurement in the x dimension is more important than the uncertainty of an individual feature point in the same dimension. The same cannot be said for the z dimension, due to the high uncertainty in the distance measurement. From these considerations, a Gaussian approximation of the form

$$\mathcal{N}(\mathbf{x}^C; \mathbf{m}, \mathbf{P}), \tag{7.11}$$

with

$$\mathbf{m} = \bar{Z}^{C,\text{2-D}}, \quad \mathbf{P} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{bmatrix}, \tag{7.12}$$

$$\sigma_x^2 = \text{cov}(Z^{C,x}), \quad \sigma_z = \frac{(\bar{Z}^{C,z})^2}{fb}\partial d, \tag{7.13}$$

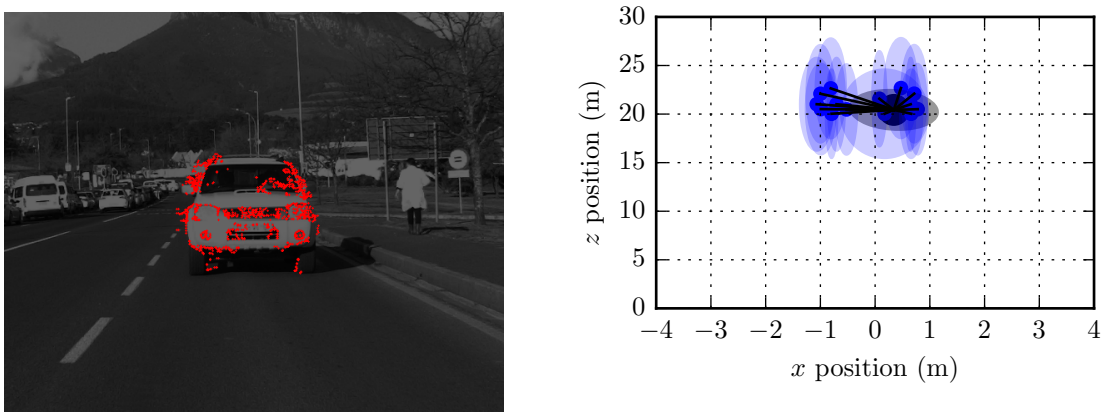
is used to approximate a vision cluster measurements Z. The notation $(\cdot)^{C,\text{2-D}}$ denotes a point in the CRF projected to two dimensions, and $\text{cov}(Z^{C,x})$ is the cluster's sample covariance in the CRF x direction.

7.3.2 Algorithm

The discussion moves on to the actual track-to-track fusion procedure. The method associates the approximated cluster measurement distributions to the mixture components of the GM-PHD radar tracker. For each cluster distribution, the nearest radar mixture component is found using the Bhattacharyya distance metric. If the resulting distance is less than a certain threshold, the cluster is grouped with the radar mixture to form a new cluster measurement Z, which is subsequently added to the fused measurement set $Z_{\text{fusion},k}$. All clusters that are output following the above steps will therefore contain a single radar track. Figure 7.3b illustrates the fusion process on the raw data shown in Figure 7.2b. The grey ellipse represents a radar mixture component, while the light blue ellipses are the Gaussian approximations of the respective vision cluster measurements. The fusion algorithm is successful in grouping the cluster measurements to radar mixture component, while the two clutter measurement clusters at $x \approx 2\,\text{m}$ (see Figure 7.2b) are correctly excluded.

The grouped cluster measurement is constructed by sampling from the radar's density, and from the vision cluster measurements that may have been associated with the particular mixture. The number of sampled points is set proportional to the components weight. If radar-vision associations exist in the group, the points sampled from the radar mixture are adjusted in height to match the spread of the vision clusters. All of the aforementioned remains in the CRF. Cluster measurements are re-projected to the WRF at the offset of the fusion routine. Three-dimensional WRF measurement clusters are therefore added to the fused set. The projection to two dimensions is only used to associate between radar mixtures and cluster measurements.

The fusion method essentially boils down to the determination of track associations, and the subsequent grouping of points from the respective sensors' tracks. Traditional track-to-track fusion approaches usually proceed in a similar manner with regard to finding the track associations. However, upon fusion, the aim is usually to refine the knowledge of a particular target using techniques such as covariance intersection [30, pp. 324–325]. The reason why the proposed algorithm deviates from the standard is due to the limitations brought forth by the particular radar used in the project. Measurements from the radar



(a) (b)

Figure 7.3: (b) Track-to-track fusion of a radar mixture component (grey ellipse) and the surrounding vision measurement clusters, which are shown by their light blue Gaussian approximations. (b) The fused cluster overlayed on the left camera image. The resulting grouping results from the application of the fusion algorithm to the ungrouped data shown in Figure 7.2b.

demonstrate low accuracy in azimuth, and the resulting state estimates are not suited to refine the knowledge of the target distributions. The radar does, however, facilitate the grouping of vision cluster measurements. Radar tracks are often centred on the surface of an object, while vision clusters follow texture-rich boundary areas. Fusing the tracks in the described manner improves the eventual grouping, especially in scenarios with high clutter levels. In view of what was just mentioned, the fusion algorithm may also be described as radar-guided clustering.

It may happen that some vision cluster measurements are not associated to any radar mixture components. A second DBSCAN clustering routine is implemented in order to group the remaining over-segmented cluster measurements. Compared to the initial feature clustering, lower density clusters are permitted in the second routine, making the implementation sensitive to surrounding clutter. However, the minimum samples constraint is increased significantly. Newly clustered groups are subsequently added to the fused measurement set $Z_{\text{fusion},k}$. Pseudo code that describes the fusion algorithm is given in Appendix A.

The track-to-track fusion algorithm is tailored for situations in which the radar is able to maintain fairly accurate state estimates of the surrounding objects in the environment. The robust sensing characteristics of radars allow such specifics, and is the reason why numerous radar-vision research efforts proceed in a similar vein. An added advantage of the proposed method is that legacy radar sensors can be used, since only its state estimates are required. Even though the fusion algorithm does not directly include the refinement of target states, subsequent inference, which is the subject of the next section, is carried out to improve the accuracy of the estimates using the fused measurements.

7.4 Extended Target Tracking

The discussion proceeds to explain the practical implementation details with regard to extended target tracking. The measurements that are output by the data fusion algorithm are in a set of sets form, which is representative of cluster measurements that originate from extended targets. These measurements are tracked using the GIW-PHD filter derived in Section 6.2, enabling both kinematic and extent information to be inferred over time. The GIW-PHD recursive equations were given in Section 6.2. The remainder of this section contains details about the models that were implemented to realise the extended target tracker.

7.4.1 Gaussian Inverse Wishart Mixture Models

The dynamic model used for the kinematic evolution of surviving targets in the extended target tracker is identical to the one introduced in Section 7.2.2, i.e. a Cartesian constant velocity model in the world reference frame (WRF). The temporal evolution of the target extent is fixed in the random matrix recursion. Consequently, Equations (6.44) and (6.45) describe the dynamic update of the extent distribution. The combined state space parameters of a single GIW mixture are given by

$$\mathbf{x} = [x, y, z, v_x, v_y, v_z]^T, \ \mathbf{X} \equiv \text{symmetric positive definite (SPD) matrix}, \tag{7.14}$$

where $\mathbf{x}$ describes the center point kinematics, and $\mathbf{X}$ is the SPD extent matrix. Recall that, in the GIW-PHD formulation, the uncertainty in these two entities are modelled by a Gaussian and inverse Wishart distribution respectively.

Observations acquired after data fusion are in the WRF, allowing a linear measurement model of the form

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix} \otimes \mathbf{I}_3, \tag{7.15}$$

$$\mathbf{R} = \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}, \tag{7.16}$$

where σ_x, σ_y, and σ_z are the standard deviations in the x, y, and z directions respectively. Mathematically, the measurement set of sets passed to Equation (6.48) at time k is of the form $Z_{\text{fusion},k} = \{Z_k^{(i)}\}_{i=1}^{N_{\text{fusion},k}}$, where $N_{\text{fusion},k}$ is the number of cluster measurements at time k, and the set elements themselves have elements of the form $[x, y, z]^T$. The measurement update is described by Equations (6.50) to (6.60).

A spontaneous birth mixture model is once more disregarded. The spawn mixture, $\upsilon_\beta(\mathbf{x}_k, \mathbf{X}_k|\cdot)$, is set up in the same vein as previously using the "unused" measurement set $Z_{\beta,k-1}$. Note that the elements of $Z_{\beta,k-1}$ are now sets instead of vectors, each describing an individual cluster measurement. Accounting for Gaussian inverse Wishart mixture components, the spawn mixture is given by

$$\upsilon_\beta(\mathbf{x}_k, \mathbf{X}_k|Z_{\beta,k-1}) = \sum_{i=1}^{N_{\beta,k-1}} w_{\beta,k}^{(i)} \mathcal{N}(\mathbf{x}_k; \mathbf{m}_{\beta,k}^{(i)}, \mathbf{P}_{\beta,k}^{(i)}) \mathcal{IW}(\mathbf{X}_k; \nu_{\beta,k}^{(i)}, \mathbf{V}_{\beta,k}^{(i)}), \tag{7.17}$$

with

$$\mathbf{m}_{\beta,k}^{(i)} = \left[\left(\bar{Z}_{\beta,k-1}^{(i)} \right)^T, 0, 0, 0 \right]^T, \tag{7.18}$$

$$\mathbf{V}_{\beta,k}^{(i)} = \tilde{Z}_{\beta,k-1}^{(i)}, \tag{7.19}$$

where $\bar{Z}_{\beta,k-1}^{(i)}$ and $\tilde{Z}_{\beta,k-1}^{(i)}$ denote the mean vector and scattering matrix of set $Z_{\beta,k-1}^{(i)}$ respectively. The corresponding uncertainties $\mathbf{P}_{\beta,k}^{(i)}$ and $\nu_{\beta,k}^{(i)}$ are again set according to a heuristic consideration.

7.4.2 Pruning and Merging

Pruning and merging in the GIW-PHD filter follows the procedure described in Section 7.2.1 for the radar tracker. However, two changes to Algorithm 1 are appropriate to account for GIW mixture components, namely the mixture-to-mixture distance metric, and the mixture reduction equations.

For the distance metric, the Bhattacharyya distance is utilised in the same manner as before, using the kinematic distribution. Granström and Orguner [71] suggests the use of the symmetric Kullback-Leibler difference (KL-diff) as a distance metric. However, practical tests showed that the corresponding threshold parameter is very difficult to tune, due to the abstract nature of GIW KL-diff distances.

The use of GIW mixture components also complicates the specification of probabilistically sound mixture reduction methods. Granström and Orguner [71] derived a solution that is based on the minimisation of the Kullback-Leibler divergence (KL-div) between the involved mixtures. The algorithm was implemented, but exchanged instead for a simpler method suggested by the same authors in earlier work [76]. It was found that the simple alternative outperformed KL-div minimisation-based mixture merging, both in terms of the quality of merged mixtures and with regard to computational performance. For a GIW state set $X_k = \{w_k^{(i)}, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)}, \nu_k^{(i)}, \mathbf{V}_k^{(i)}\}_{i=1}^{J_k}$, and given that the indices in the set L should be merged, the reduction equations are given by

$$\bar{w}_k = \sum_{i \in L} w_k^{(i)}, \tag{7.20}$$

$$\bar{\nu}_k = \frac{1}{\bar{w}_k} \sum_{i \in L} w_k^{(i)} \nu_k^{(i)}, \quad \bar{\mathbf{V}}_k = \frac{1}{\bar{w}_k} \sum_{i \in L} w_k^{(i)} \mathbf{V}_k^{(i)}. \tag{7.21}$$

The equations for the merged kinematic mean and covariance are unchanged from those given in Algorithm 1.

GIW-PHD filtering marks the final step of the radar-vision data fusion framework. The resulting state estimates represent the output of the data fusion system. As an example, the extended target track that follows Figures 7.2a and 7.3b is shown in Figure 7.4.

Figure 7.4: Visualisation of a target's extent estimate that result from GIW-PHD filtering.

Chapter 8

Results

In this chapter, the results of the data fusion system are be presented. The evaluation procedure involves computer simulations as well as tests on real-world data. To begin with, the metrics that allow the quantitative representation of the system's performance are discussed. Thereafter follows the evaluation procedure and a presentation of the results for both the simulation and real-world evaluation setups.

8.1 Multi-Target Tracking Performance Metrics

In order to compare the system's behaviour to similar methods, it is necessary to extract quantitative performance metrics. The multiple object tracking metrics proposed by Bernardin and Stiefelhagen [79], along with the optimal subpattern assignment (OSPA) error of Schuhmacher et al. [80] are adopted for the presentation of the results.

8.1.1 Multiple Object Tracking Precision and Accuracy

In the work of Bernardin and Stiefelhagen [79], tracking performance is quantified from the validity of correspondences and the consistency of tracking over time. A correspondence between a ground truth object and tracker hypothesis is labelled invalid if the distance between them exceeds a certain threshold. A one-to-one matching procedure of ground truth object and tracker hypothesis pairs that pass the threshold test are used to find valid correspondences. The measure of consistency pertains to the tracker's ability to correctly label the identity of an object as it is tracked over time. Based on the aforementioned, Bernardin and Stiefelhagen define two metrics to quantify multi-target tracking (MTT) performance, namely multiple object tracking precision (MOTP)

$$\text{MOTP} = \frac{\sum_{i,k} d_k^i}{\sum_k c_k}, \tag{8.1}$$

and multiple object tracking accuracy (MOTA)

$$\text{MOTA} = 1 - \frac{\sum_k (m_k + fp_k + mme_k)}{\sum_k g_k}, \tag{8.2}$$

where

- d_k^i is the distance between the i^{th} ground truth object and its corresponding tracker hypothesis at time k;

- c_k is the number of matches found at time k;
- m_k is the number of missed detections at time k;
- fp_k is the number of false positives at time k;
- mme_k is the number of mismatches (identity switches) at time k; and
- g_k is the number of objects present at time k.

MOTP specifies the misalignment between ground truth and predicted bounding boxes, whereas MOTA indicates the ratio of tracker errors to the number of ground truth objects. If the distance measure is exchanged for an overlap measure, then a value of unity is desirable for both MOTP and MOTA. The lower performance bound of MOTP coincides with the minimum overlap that is still considered a valid hypothesis. MOTA values may drop to below zero when the tracker is more prone to make a mistake than a correct hypothesis. The evaluation that is to follow will use overlap distance measures to quantify the performance of object extent estimation. Moreover, it should be noted that the metrics defined in Equations (8.1) and (8.2) are popular in computer vision research, where overlap measures are mostly used.

8.1.2 Optimal Subpattern Assignment

The MTT research community generally adopts different metrics than those encountered in computer vision. Concepts such as false positives and missed detections are more commonly associated with object classification, whereas MTT evaluation place focus on the correct estimation of the number of targets (cardinality) and localisation accuracy.

An accepted method for MTT evaluation is the OSPA metric devised by Schuhmacher et al. [80]. The algorithm finds the optimal correspondence solution of tracker hypotheses to ground truth objects using a distance function, before calculating an error score based on the correspondences. Given the tracker hypotheses set $X = \{\mathbf{x}^{(i)}\}_{i=1}^{m}$ and the ground truth set $Y = \{\mathbf{y}^{(i)}\}_{i=1}^{n}$, the OSPA error is of the form

$$\text{OSPA}(X, Y) = \left(\frac{1}{n} \left(\sum_{i=1}^{m} \left(d^c(\mathbf{x}^{(i)}, \mathbf{y}^{(\pi(i))}) \right)^p + c^p(n-m) \right) \right)^{1/p}, \tag{8.3}$$

where $d^c(\cdot, \cdot)$ is a distance function accounting for localisation errors, and the second term in the summation accounts for cardinality errors. The constant c determines the maximum value that the distance function may return, i.e. $d^c(\cdot, \cdot) \in [0, c]$. The notation $\mathbf{y}^{(\pi(i))}$ is used to indicate the optimal ground truth correspondence of hypothesis $\mathbf{x}^{(i)}$. If $m > n$, the error is given by $\text{OSPA}(Y, X)$.

Two parameters are adjustable in Equation (8.3), namely the order p and the cut-off constant c. Intuitively, the OSPA metric can be interpreted as a p^{th} order per-object error [80]. The cut-off parameter is used to limit the penalty that a correspondence or cardinality error contributes. Small values of c pronounces localisation errors while lessening the effect of cardinality errors. A high cut-off value has the opposite effect.

In order to provide results that are applicable to both the MTT and computer vision communities, the metrics that have been discussed in this section will be applied according to the related field being evaluated. In particular, the OSPA calculations will be formulated

in such a way so as to represent the error in the estimate of a target's centre point, while MOTP and MOTA will use overlap measures to describe the performance of extent estimation. The abovementioned use of the OSPA metric is consistent with the general use thereof in target tracking research, while the proposed use of the MOT metrics comply with computer vision tendencies.

8.2 Simulation Setup

A multi-target tracking scenario was simulated in order to test various aspects of the proposed data fusion system. A two-dimensional space is considered in the simulation and the sensor frame is assumed to remain stationary. The scenario includes three targets moving in a straight line, clutter interference, and crossing tracks toward the end of the simulation. A graphical representation of the simulation is shown in Figure 8.1. The plot contains the accumulated measurements over the 100 time steps comprising the simulation. The apparent lines represent target tracks, and are caused by the aggregation of target observations. All measurements are plotted with a degree of transparency except for those that occur at the last time step, making it clear that the targets move from left to right. Generic, unspecified units are used in the simulation.

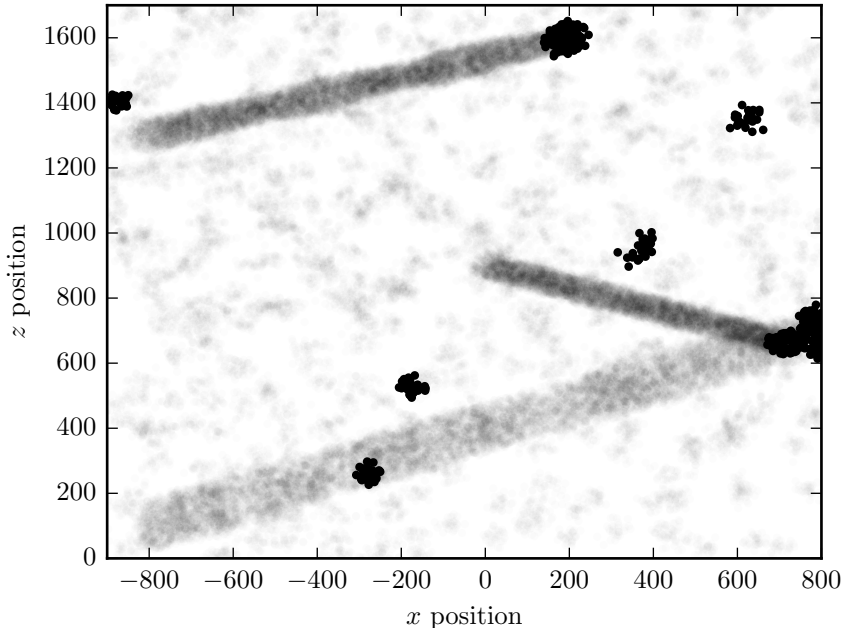


Figure 8.1: Plot of the accumulated target and clutter measurements over the 100 time step simulation. All measurements are plotted with a degree of transparency, except for those that occur at the final time step. The apparent lines represent target tracks, and are caused by the aggregation of target observations.

The true position of the i^{th} target's centre over time is given by

$$\mathbf{x}_k^{(i)} = [x_k^{(i)}, z_k^{(i)}]^T = [x_0^{(i)}, z_0^{(i)}]^T + [v_x^{(i)}k, v_z^{(i)}k]^T, \tag{8.4}$$

where $[x_0^{(i)}, z_0^{(i)}]^T$ is the starting position, and $v_x^{(i)}$ and $v_y^{(i)}$ are the velocities in the x and z directions respectively. The true centre point position is therefore fully described by

the parameter set $\{x_0^{(i)}, z_0^{(i)}, v_x^{(i)}, v_z^{(i)}\}$. For the targets in Figure 8.1, these sets are given by $\{-800, 100, 16, 6\}$, $\{0, 900, 7.2, -2.4\}$, and $\{-800, 1300, 10, 3\}$, respectively. The true target extent is given by an ellipse, parametrised by the length of its major and minor axes, and its orientation angle. Some or all of each ellipse's extent parameters are set to vary sinusoidally in time.

Target measurements are assumed to be uniformly distributed over the target's extent ellipse. To generate such measurements, a rejection sampling strategy is implemented[1]. The number of measurements is set proportional to the target's area. Observations are generated in the Cartesian format $[x, z]^T$. Measurement generation is concluded by adding normally distributed noise to the individual measurements that result from rejection sampling.

Clutter measurement generation is performed in two steps. First, the centre point of a clutter measurement is sampled uniformly from the interval representing the surveillance area, which coincides with the limits shown in Figure 8.1. To generate an extended target cluster measurement, the centre point is subsequently used to sample from a bivariate normal distribution with a random covariance matrix. Specifically, for each new clutter measurement, the covariance matrix is sampled from an inverse Wishart distribution with a low degrees of freedom, i.e. high uncertainty. The resulting cluster measurements demonstrate a high degree of randomness, as would be expected from clutter.

The simulated measurements at each time steps is of the form

$$Z_k = \{Z_k^{(i)}\}_{i=1}^{N_{\text{sim},k}}, \tag{8.5}$$

where $N_{\text{sim},k}$ is equal to the number of target-originating cluster measurements plus the number of clutter cluster measurement at time k. Five clutter measurements were added at each time step. Note that the representation given in Equation (8.5) implies perfect clustering.

8.3 GIW vs Gaussian Measurement Model

Mention was made in Section 2.4 and Chapter 6 about the importance of appropriate measurement models when objects generate numerous spatially distributed measurements per scan. To provide quantitative results of the point vs extended target assumption, the evaluation will set of with a comparative analysis of the GM-PHD and GIW-PHD filters in an extended target tracking scenario. The analysis is based on the simulation environment described in the previous section.

8.3.1 PHD Mixture Models

Testing the filters require the implementation of appropriate mixture models for the scenario under consideration. The models are chosen in accordance with the ones used in the data fusion algorithm of Chapter 7. An obvious difference is a lower dimensional state space, due to the fact that the environment is two-dimensional. For the GM-PHD filter, the state vector of a single mixture component is given by

$$\mathbf{x} = [x, z, v_x, v_z]^T. \tag{8.6}$$

[1]Rejection sampling is explained in Appendix B.

The state vector in Equation (8.6) is also used to model the centre point kinematics for the GIW-PHD filter. The kinematic state in both filters is assumed to evolve according to the constant velocity model [63]. A linear measurement model of the form

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \end{bmatrix} \otimes \mathbf{I}_2, \tag{8.7}$$

$$\mathbf{R} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{bmatrix}, \tag{8.8}$$

is applied in the kinematic update equations of both filters. Updating the extent distribution of the GIW-PHD filter is done in the same manner as before (see Section 7.4.1), and the spawn mixture again uses the "unused" measurement set.

8.3.2 Simulation Results

This section contains the results of the application of the filters described in the previous section to the simulation environment of Section 8.2. To recap, the evaluation considers an extended target environment. Hence, the simulated measurements at each time step are passed to the respective filters as follows: The measurement set of Equation (8.5) is passed as-is to the extended target GIW-PHD filter, with perfectly clustered set elements. In contrast, the point target GM-PHD filter is applied naively to the measurement set with the assumption that each target will generate at most one measurements per time step. In other words, the set elements of the measurement set are unpacked, and the resulting set of vectors passed to the GM-PHD filter. This procedure may seem biased toward the GIW-PHD filter. However, the aim here is to evaluate the neglect versus application of extended target models in extended target tracking scenarios (Granström et al. [81] present similar results for their GIW-PHD filter.). Note that 5 % of all target related measurements are randomly discarded to simulate missed detections.

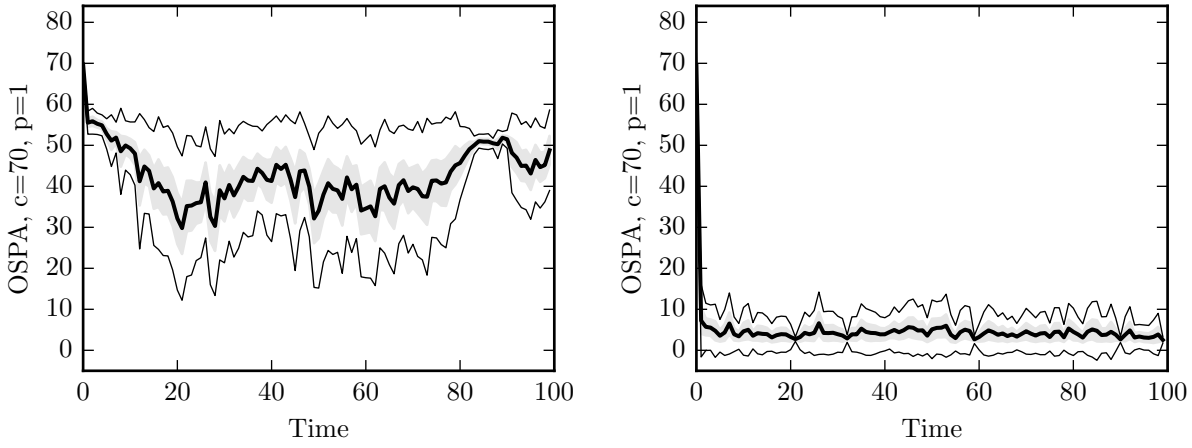
The simulation results with regard to localisation accuracy are given by the OSPA error vs time plots shown in Figures 8.2a and 8.2b. The order parameter is set to 1, while Euclidean distance between the ground truth position and kinematic mean is used as the error function. Using a first order error means that the OSPA values shown in the figures can be interpreted as a variant of Euclidean distance. The cut-off parameter is set according to the recommendations of the authors who derived the OSPA metric[2].

Figures 8.2a and 8.2b represent the result of 30 Monte Carlo repetitions. The thick black line represents the mean Monte Carlo result. The thin black line corresponds to the $\pm 1\sigma$ error bounds, while the shaded area represents the 95 % confidence interval[3]. It is clear from the figures that the GIW-PHD filter significantly outperforms the GM-PHD filter in terms of localisation accuracy. The false assumption that targets generate only a single measurement per time step proves detrimental to the performance of the GM-PHD filter.

Figures 8.3a and 8.3b shows the cardinality results of the Monte Carlo simulation. The solid lines represent the ground truth number of targets, while the Monte Carlo mean is given by the dashed line. The GM-PHD filter is completely unable to correctly estimate the number of targets, while the GIW-PHD maintains a small error margin throughout

[2]The cut-off value is considered small if it corresponds to typical localisation errors, and it is considered large if it corresponds to the maximum distance between objects [80]. Considering the object separation and surveillance area, a value of 70 was deemed fit.

[3]The estimated 95 % confidence bounds are given by $\pm 1.96\sigma/\sqrt{N_{mc}}$, where σ is the standard deviation of the data, and N_{mc} is the number of Monte Carlo repetitions.



(a) (b)

Figure 8.2: Monte Carlo simulation results of (a) the GM-PHD filter and (b) the GIW-PHD filter applied to the scenario of Section 8.2. The mean OSPAs are shown by the thick black lines. The thin black lines represent the $\pm 1\sigma$ bounds, while the shaded areas represent the 95 % confidence interval. The violation of the point target assumption impedes the performance of the GM-PHD filter, while the GIW-PHD maintains accurate localisation throughout the simulation.

the simulation. The cardinality error of the GM-PHD filter is the main contributor to its high OSPA error.

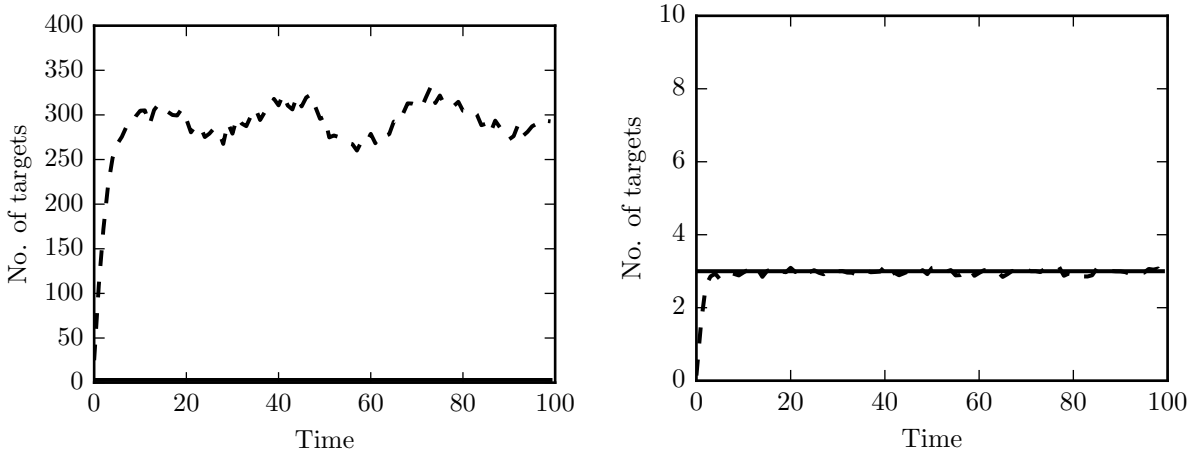
Recall that the estimated cardinality is the sum of weights in the target state set, and not the number of elements therein. For the calculation of the OSPA error, the number of elements in the state set constitutes the parameter that represents the estimated number of targets. The OSPA error of the GM-PHD filter would have saturated to the cut-off value had the cardinality estimate been used. It is therefore clear that the number of targets in the state set is less than the estimated cardinality, implying that the weights of the individual mixture components exceed 1. The high mixture weight is a consequence of the violation of the point target assumption.

The results show that the point target assumption leads to extremely poor estimation performance when targets generate numerous measurements per scan. The inability to correctly estimate the number of targets degrades the filter's accuracy to such an extent that the resulting state estimates are completely untrustworthy.

The GIW-PHD filter provides reliable state estimates, remaining largely unaffected by clutter interference and missed detections. Moreover, the adaptive spawn mixture strategy is shown to react quickly when the target measurements appear.

8.4 Data Fusion Simulation

The previous section established the importance of extended target measurement models in environments that require such models. It also demonstrated the localisation accuracy of the random matrix model for extended target tracking. In this section, the simulation results regarding the data fusion algorithm developed in this project will be presented.



(a) (b)

Figure 8.3: Monte Carlo simulation results of (a) the GM-PHD filter and (b) the GIW-PHD filter applied to the extended target scenario of Section 8.2. The ground truth cardinality is shown by the solid lines, while the dashed lines represents the respective mean cardinality estimates. Violation of the point target assumption causes the GM-PHD filter to significantly overestimate of the number of targets.

8.4.1 Simulation Modifications and PHD Models

The analysis is based on a slightly altered simulation environment from the one described in Section 8.2. No alterations is made with regard to true target positions and their extent. The only changes that are applied relate to the generation and passing of the simulated measurements to the respective filters and the use of a non-linear measurement model in the GM-PHD filter. The resulting GM-PHD filter represents the radar state estimator of the fusion algorithm. The GIW-PHD filter represents the output extended target tracker of the data fusion algorithm, which was described in Section 7.4.

The adapted filter will first be presented. In keeping with the GM-PHD filter that is implemented for radar-based tracking in the data fusion algorithm, a non-linear measurement model is included. As before, the only changes from the implementation described in Section 7.2 relate to the dimensionality reduction. The dynamic model is identical to the one used in the simulation of Section 8.3. The measurement model assumes range bearing measurements, and is of the form

$$h : \mathbf{x} \rightarrow \mathcal{N}(\mathbf{z}; \bar{\mathbf{z}}, \mathbf{R}), \tag{8.9}$$

with $\mathbf{x}$ as in Equation (8.6), and

$$\mathbf{z} = [r, \alpha]^T, \tag{8.10}$$

$$\mathbf{R} = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\alpha^2 \end{bmatrix}. \tag{8.11}$$

The unscented transform is once again used to approximate the non-linear transformation of Gaussian mixture components.

Moving on to the changes in measurement handling, the measurement generation process is performed separately for the GM-PHD and GIW-PHD filters. At each time step,

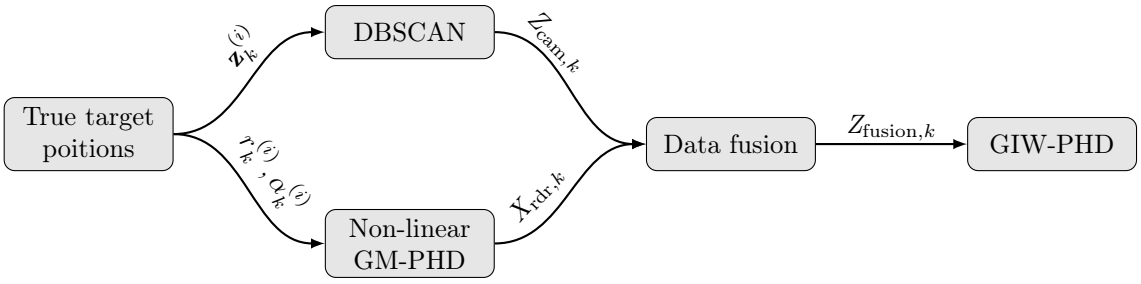


Figure 8.4: Diagram of the track-to-track fusion simulation. The processing pipeline mimics the actual data fusion algorithm of Section 7.3.

the mean position of a target is converted to polar coordinates and corrupted with normally distributed range-bearing noise, before being passed to the non-linear tracker. The means of clutter measurements, in polar coordinates, are also included. Cluster measurements are generated in exactly the same manner as before. However, ideal clustering is done away with in order to realistically simulate the data fusion algorithm. Therefore, true target and clutter measurement clusters are sent as ungrouped vectors to the clustering block. The processing pipeline mimics the actual data fusion algorithm of Section 7.3, and is shown in Figure 8.4. Hence, the same fusion algorithm terminology will be used in order to emphasise the context, e.g. vision cluster measurements and radar mixture components. Note that as before, 5 % of target related measurements are randomly discarded before being passed to the respective filters.

For fusion simulations, the multiple object tracking precision and accuracy are also included. The area intersection over union of ground truth and GIW-PHD extent ellipses constitute the overlap metric. The intersecting area of two ellipses is calculated by means of a polygon approximation of the ellipses' boundaries, since no closed form expression exist that give the intersecting area of general ellipses. The overlap threshold is set to 0.3. Small position and orientation mismatches may induce large overlap mismatch errors for two ellipses, motivating the relatively low overlap requirement.

MOTP and MOTA results are usually calculated from an entire sequence. If a parameter exist that directly influences the algorithm's detection performance, MOTP and MOTA are often plotted against different values of the particular parameter to determine the optimal value thereof. The process is reminiscent of precision-recall curves. Parameter variation, however, is not included for the evaluation, since there are multiple parameters that influence the estimation performance. Instead, the Monte Carlo average of MOTP and MOTA will be given.

8.4.2 Track-to-Track Fusion

The results of applying the track fusion algorithm to the simulation environment described in the previous section will now be presented. The localisation accuracy, in terms if the OSPA error, for the eventual state estimates that are output by the GIW-PHD filter is shown in Figure 8.5. It is apparent that the error is consistently higher than what was achieved in the perfectly clustering scenario of Section 8.3.2. However, the error remains fairly low considering the noise levels and number of clutter measurements. A marked performance deterioration occurs after the 80^{th} time step when two of the targets near each other. The reason for the increased OSPA error can be deducted from the corresponding cardinality results shown in Figure 8.6, which suggests that the targets are perceived as one

somewhere in the processing chain. In this case, the reason for the error is due to the initial DBSCAN clustering routine that groups the overlapping measurements into one cluster measurement. A similar event can occur if closely spaced radar mixture components are grouped during the merging procedure. Adjacent vision cluster measurements will then most likely be associated to the grouped mixture, again resulting in a single fused cluster measurement. The merging of closely spaced targets is referred to as track coalescence; a property inherent to soft association tracking algorithms, including the GM-PHD filter.

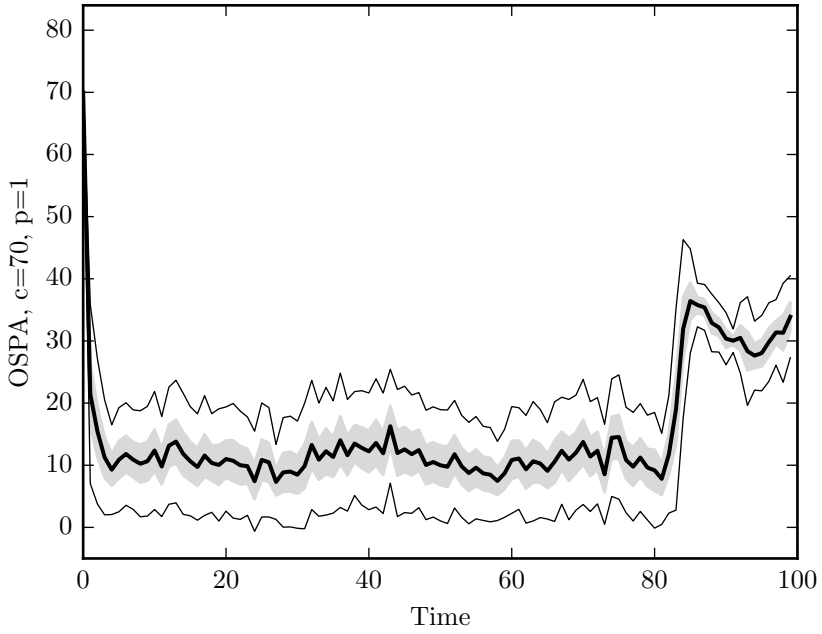


Figure 8.5: Monte Carlo OSPA simulation results of the track fusion algorithm applied to the scenario of Section 8.4.1.

During the track fusion simulation, the average MOTP and MOTA amounted to 0.48 and 0.66 respectively. Given the time varying extent parameters, the MOTP indicates satisfactory performance with regard to the extent precision of tracker hypotheses that were matched to ground truth targets. Factors that impeded the accuracy were largely due to false hypotheses and identity switches caused by clutter.

The localisation error of the radar's non-linear GM-PHD filter is not shown here. However, the cardinality shown in Figure 8.6 confirms that it performed satisfactory. The filter proves effective in maintaining reliable state estimates in the presence of clutter and missed detections. Also, the adaptive spawn mixture strategy is again very quick to respond to the advent of target measurements.

8.4.3 No Fusion

Results of an alternative processing pipeline that disregards radar information will be presented in this section. The track fusion simulation setup is adopted, but no measurements are passed to the GM-PHD filter that resembles the radar. The goal is to evaluate whether data fusion improves the eventual estimation performance.

The results of no data fusion is shown in Figures 8.7 and 8.8. Note that track coalescence occurs at an earlier time step than in the fusion simulation. This proves that the

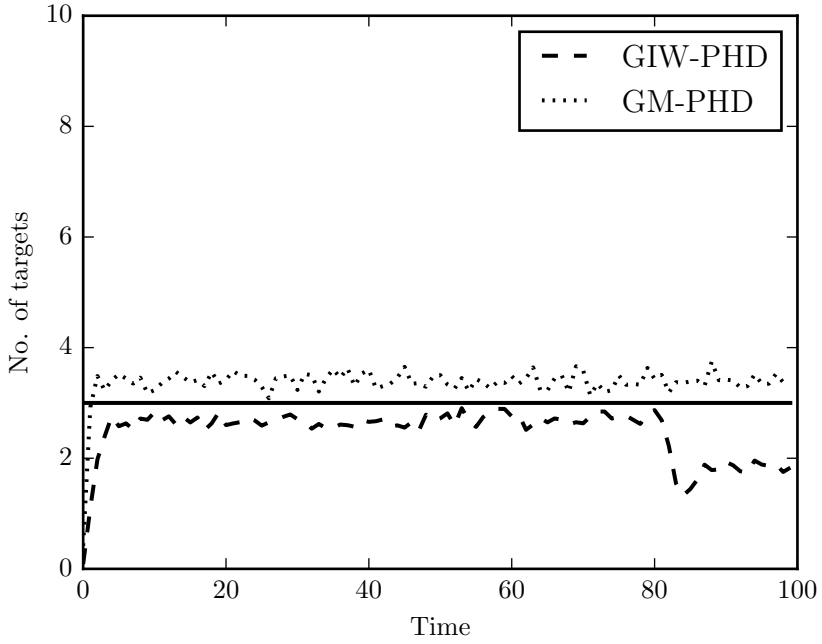


Figure 8.6: Monte Carlo cardinality simulation results of the track fusion algorithm applied to the scenario of Section 8.4.1. The GM-PHD filter maintains an accurate estimate of the number of targets throughout the simulation. DBSCAN clustering, however, fails to separate the measurements generated from the crossing tracks near the 80^{th} time step, leading to a reduced estimate in the number of targets for the GIW-PHD filter.

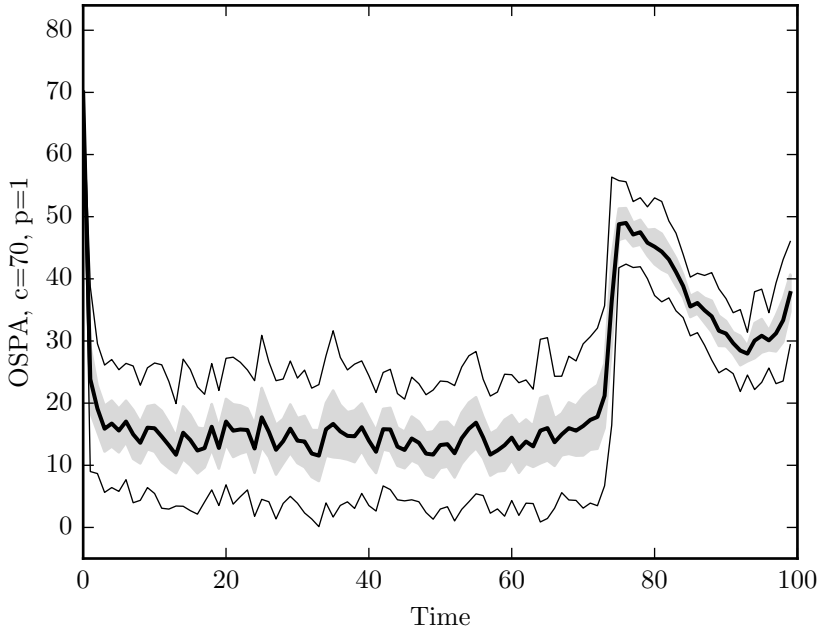


Figure 8.7: Monte Carlo OSPA results for the simulation environment of Section 8.4.1, but without radar information. The second clustering routines is more sensitive to adjacent clutter and crossing tracks, leading to a decrease in localisation accuracy.

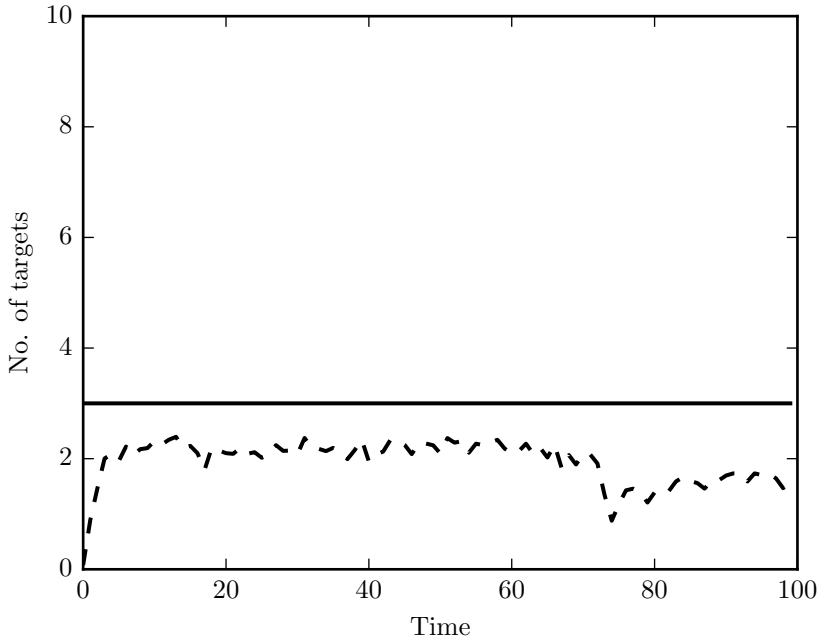


Figure 8.8: Cardinality simulation results of the single-sensor simulation.

association of cluster measurements to radar mixture components assists the formation of reliable clusters. Without track-to-track fusion, all the initial DBSCAN cluster measurements proceed to the second DBSCAN routine, in which less dense clusters are permitted. Consequently, adjacent clusters merge sooner. Clustering with the relaxed requirement makes the vision-only method sensitive to clutter interference.

The cardinality and OSPA errors preceding the coalescence event are consistently higher than the corresponding results of the track fusion method. The intermittent merging of target and clutter clusters impedes the quality of the measurements passed to the GIW-PHD filter. The cardinality plot of Figure 8.8 shows a persistent under-estimation of the number of targets, suggesting that the lower quality measurements cause occasional track loss.

MOTP and MOTA also demonstrate decreased performance, with values of 0.46 and 0.41 respectively. MOTP is only marginally lower, and can be contributed to the degraded quality of the cluster measurements. The drop in accuracy is largely due to an increase in targets being lossed. The cardinality plot supports this argument.

The simulation shows that the track-to track fusion method increases the system's performance in the presence of clutter. It also highlights the sensitivity of the random matrix model to imperfect clustering. A consistent error increase is seen compared to the earlier extended target simulation of Section 8.3 which received perfect clusters.

8.5 Practical Results

The discussion now advances to the presentation of the system's performance on real-world data. In order to conduct the evaluation, datasets were gathered during regular highway driving using the hardware described in Section 3.1. The robot was fastened to

the back of a pick-up truck, with the sensors facing rearwards. The analysis will resemble that of the fusion simulation of Section 8.4.

Three real-world sequences, each comprising 100 frames, are used for comparative testing of various entities of the data fusion system. The track fusion processing pipeline described in Section 7.3 is applied as-is to these sequences. Comparative results are also extracted for the scenario in which radar data is discarded in order to evaluate the data fusion's impact on the overall performance gain or loss. Furthermore, the results of state estimation are compared against raw measurements to assess the value of tracking. The sequences will be referred to by the respective locations where they gathered in and around Stellenbosch, namely Helshoogte, R44, and Merriman. The Helshoogte sequence was used as a training dataset to tune the various clustering and PHD filter parameters.

8.5.1 Ground Truth Labelling

Ground truth information is required for the calculation of performance metrics. The nature of the data collection platform and the environments considered, however, renders such information unavailable. It is common practice to resort to distance measures in the image plane if the dataset contains images, since ground truth target positions can easily be extracted by means of manual labelling. This is also the procedure adopted here for the practical evaluation. The VATIC annotation tool was used to facilitate ground truth labelling [82].

Ground truth labels takes the form of bounding boxes in the image plane. To obtain compatible tracker hypotheses, the unscented transform is used to project WRF state estimates, as output by the GIW-PHD filter, to the image plane. An upright bounding box is then fitted to the resulting two-dimensional extent ellipse to represent the tracker's hypothesis.

As before, the OSPA metric describes the error in the estimate of a target's centre point, while MOTP and MOTA describe extent estimation performance. The OSPA error is now in units of pixels, and bounding box intersection over union is used as an overlap measure. The overlap threshold is increased to 0.5, as is common when using bounding boxes [83]. Traditional precision and recall metrics are also included, defined as

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \tag{8.12}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False Negatives}}. \tag{8.13}$$

Recall represents the percentage of ground truth objects that are successfully identified, whereas precision gives the percentage of hypotheses that are correct. The same overlap criterion is used to classify a hypothesis as valid or not.

8.5.2 Helshoogte Sequence

The evaluation results of the training sequence will first be presented. A still frame from the sequence is shown in Figure 8.9. The sequence contains a single moving vehicle driving behind the sensing platform. Furthermore, the relative arrangement between the sensor platform and the vehicle remains similar to the scenario shown in Figure 8.9 for the duration of the dataset.

Figure 8.10 shows results for various of the output stages of the data fusion system. The term 'fusion' is applied to scenarios where radar information is fused with the vision

Figure 8.9: Frame from the Helshoogte dataset. The sequence contains a single moving vehicle driving behind the sensing platform.

cluster measurements as described in Section 7.3, while the term 'no fusion' or 'vision-only' refers to the same processing pipeline, but without any radar information. The subfigures portray the following: The top left figure shows the GIW-PHD filter output OSPA errors of the fusion and vision-only methods, while the top right figure illustrates the corresponding cardinality estimates. The bottom left figure compares the fusion method's GIW-PHD filter output to the measurements passed to it. Finally, the bottom right figure shows the OSPA errors of the output cluster measurements that result from the fusion and no fusion processing chains. The multi-target state set X_k is used to label results that pertain to filter hypotheses, while the measurement set symbol Z_k is used to label results that pertain to measurements.

Figure 8.10a illustrates the OSPA error of the GIW-PHD filter hypotheses over the dataset sequence for the fusion and no fusion scenarios respectively. The disregard of radar measurements leads to a small increase in localisation error, amounting to an average difference of 1.4 pixels, which is trivial. The reason for the similar performance is clear from Figure 8.10d, which shows the OSPA errors of the measurement clusters passed to the GIW-PHD filter for the respective scenarios. The fusion and no fusion methods demonstrate very similar localisation errors, suggesting comparable performance for this particular sequence. Interesting to note is that the areas of low OSPA errors in Figure 8.10a correspond to time steps where the cardinality was estimated correctly. Here, the OSPA error is of the order of a few pixels, indicating accurate centre point estimation for persistent tracks. However, clutter objects were often tracked, causing the occasional spikes in the OSPA error. Note that this is a consequence of the measurement extraction procedure and not the tracking as such.

Figure 8.10c shows the interesting comparison of the GIW-PHD filter's OSPA error, along with the corresponding error of the measurement set passed to the filter. Both result from the track fusion processing procedure. The mean OSPA errors are 34.1 and 58.5 for the filter and measurements respectively. These numbers substantiate the belief that state estimation should improve perception performance. Note that a bounding box is fit to cluster measurements in order to calculate overlap and distance measures for the clusters. The bounding boxes then serve as the clusters' hypotheses, in the same manner

that target extent bounding boxes represent the GIW-PHD filter hypotheses.

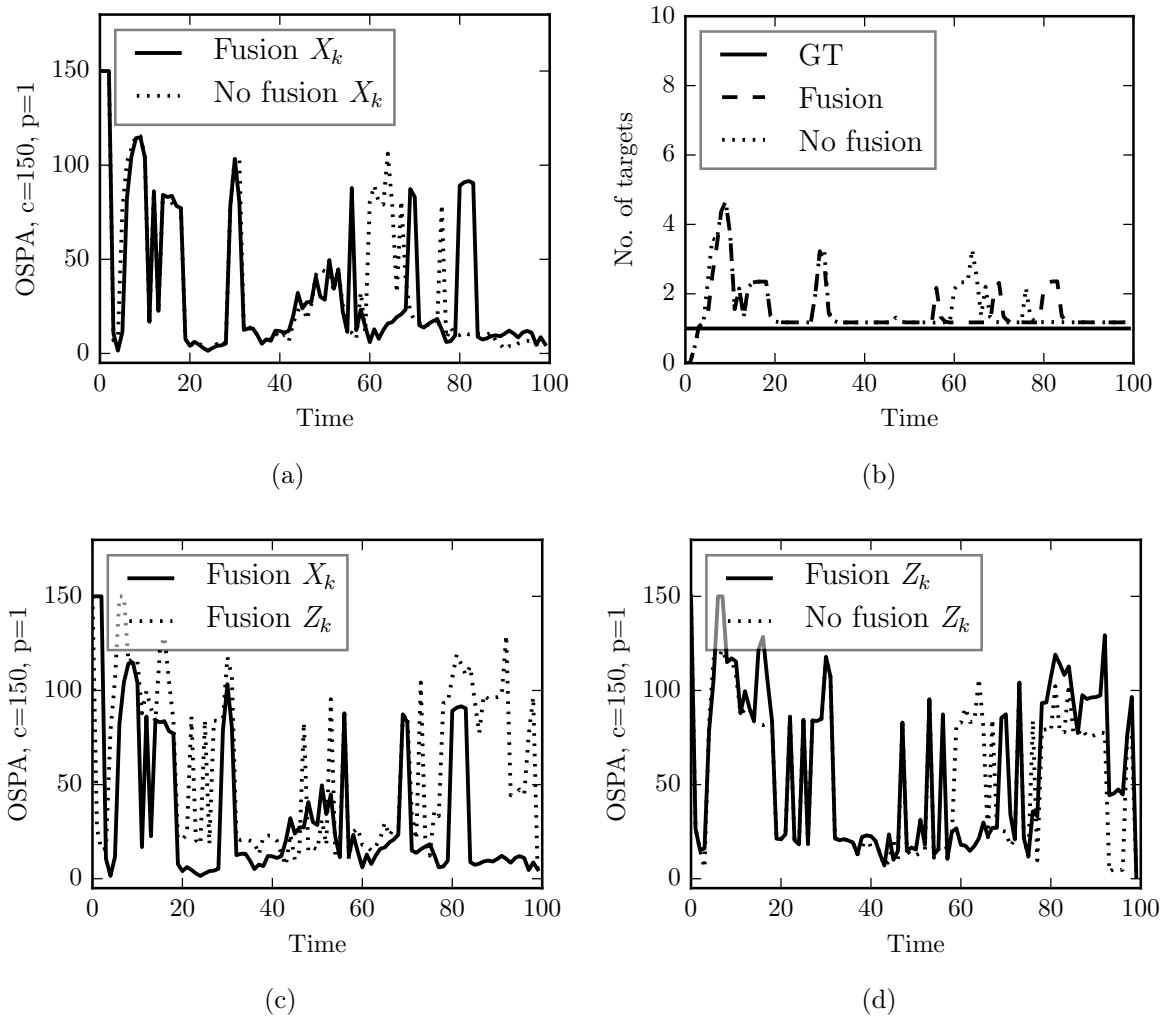


Figure 8.10: OSPA and cardinality evaluation results for the Helshoogte sequence. (a) OSPA error of the GIW-PHD filter hypotheses for the track fusion algorithm and for vision-only processing. (b) The corresponding cardinality results. (c) OSPA error of both the GIW-PHD filter hypotheses and the accompanying fused cluster measurement set. (d) OSPA error of the cluster measurement set passed to the GIW-PHD filter for both the fusion and no fusion scenarios.

Table 8.1: Tracking performance metrics for the Helshoogte sequence. The fusion and vision-only processing methods demonstrate similar performance. Estimation generally leads to improved results compared to raw measurements, except for MOTP.

	MOTP	MOTA	ID switches	Precision	Recall	$\overline{\text{OSPA}}$
GIW-PHD fusion	0.75	0.63	2	0.76	0.95	34.1
GIW-PHD no fusion	0.75	0.57	3	0.73	0.95	35.5
$Z_{\text{fusion},k}$ fusion	0.71	-0.07	N/A	0.48	0.79	58.5
$Z_{\text{fusion},k}$ no fusion	0.79	-0.19	N/A	0.44	0.75	54.6

To evaluate the algorithm's performance with regard to extent estimation, the applicable moving object tracking metrics for the sequence are also extracted. Table 8.1 contains the relevant results for the various entities comprising the system. A similar pattern can be seen as before with regard to the fusion and no fusion performance. The MOTP measures show little variation, indicating that extent estimation is consistent for valid hypotheses. Valid hypotheses of the camera-only measurements outperform those that result from state estimation, due to the increased area of the bounding boxes that are fit to target ellipses. Fused cluster measurements are expected to follow this trend. However, the MOTP result of 0.71 indicate the contrary. The likely reason is the following: A radar track may drift off of an object, but vision cluster measurements may still be associated to the particular mixture, especially if the track's uncertainty is high. Resulting cluster measurements then tend to be larger than the actual size of an object. The lower MOTP score is due to the reduced intersection over union value that follow the larger cluster measurements.

Analysis of the MOTA values again reinforce the need for state estimation, as can be seen in Table 8.1. Negative values result from the respective techniques' cluster measurement evaluation, indicating that the hypotheses contain more errors than the number of objects in the sequence. False positives are the main contributor to the low accuracy. The clutter modelling of the GIW-PHD filter helps alleviate the effect of false measurements in the eventual filter hypotheses.

8.5.3 R44 Sequence

The evaluation results of the R44 dataset will now be presented. The dataset was not used for parameter tuning, contrary to the Helshoogte sequence. A still frame from the dataset is shown in Figure 8.11. The R44 sequence is similar to the Helshoogte sequence. It again contains a vehicle that trails the sensing platform. A pedestrian also appears near the end of the sequence.

Figure 8.11: Frame from the R44 dataset. The sequence contains a moving vehicle driving behind the sensing platform, and a pedestrian appears near the end of the sequence.

The centre point OSPA error plots are shown in Figure 8.12. For the most part of the sequence, and for both the data fusion and camera-only scenarios, the GIW-PHD filter maintains a localisation error of roughly 20 pixels. This error is quite small considering that the image dimensions are 1280×960 pixels. The error increases near the end of the simulation, when the system is unable to detect a pedestrian. Hypotheses that result from track-to-track fusion are again more accurate than their un-fused counterparts, due to the slightly increased quality of measurements that result from fusion processing (see Figure 8.12d). Cluster measurements that are output without the aid of radar data deteriorates when the pedestrian appears near the 80^{th} time step. Even though the fusion pipeline also miss detects the pedestrian, the other target remains detected due to the influence of its radar track. The estimated cardinality shown in Figure 8.12b confirms the aforementioned event.

Figure 8.12c shows the OSPA error of the GIW-PHD filter hypotheses alongside the OSPA error of the hypotheses extracted from the cluster measurements for the track fusion scenario. A notable performance increased is brought about by the extended target tracker, with the mean error being decreased from 47.8 to 33.8. Table 8.2 also indicates the performance increase for the case where only camera data is used.

Table 8.2: Tracking performance metrics for the R44 sequence. The OSPA error values indicate that GIW-PHD filtering leads to a notable increase in localisation accuracy. The MOTA scores also show that filtering leads to fewer overall tracking errors. Valid measurement clusters show improved overlap compared to GIW-PHD filter extent estimates, as shown by the MOTP values.

	MOTP	MOTA	ID switches	Precision	Recall	$\overline{\text{OSPA}}$
GIW-PHD fusion	0.59	0.50	0	0.81	0.65	33.8
GIW-PHD no fusion	0.61	0.61	0	0.89	0.70	38.0
$Z_{\text{fusion},k}$ fusion	0.76	0.33	N/A	0.66	0.69	47.8
$Z_{\text{fusion},k}$ no fusion	0.76	0.33	N/A	0.66	0.68	49.4

The metrics describing the extent evaluation are given in Table 8.2. The MOTP values indicate a consistent decrease in the extent accuracy when comparing the filter estimates to the cluster measurements. As before, the main reason is due to the increased area that result from fitting a bounding box to an ellipse. The MOTA values indicate that data fusion resulted in a drop in accuracy concerning the GIW-PHD hypotheses. This can again be explained by poor radar state estimates that drift away from the true target position. In such scenarios, the fused cluster measurements may be larger than the actual target, and the GIW-PHD extent estimate will grow accordingly. The larger measurement may still exhibit the required 50 % intersection over union criteria to be classified as a valid detection. However, the filter hypothesis contains added area after bounding box fitting, which can reduce the overlap score such that the hypothesis is labelled invalid. Hence the reduced MOTA value.

8.5.4 Merriman Sequence

The final practical results relate to the Merriman sequence. The scene provides evidence of some of the shortcomings of the data fusion system. Figure 8.13 shows two still frames from the dataset with the raw feature detections overlaid. The scenario is again fairly

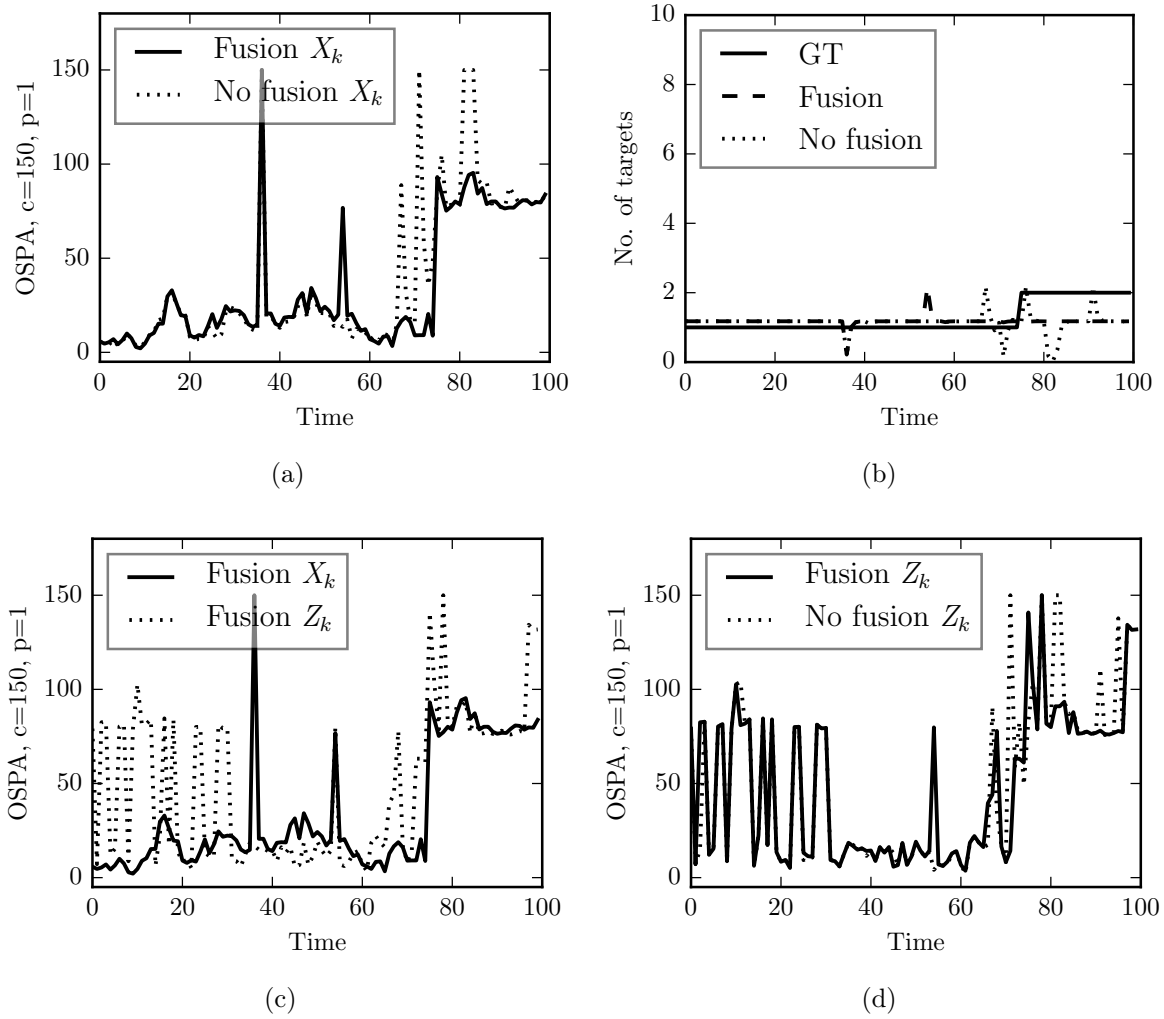
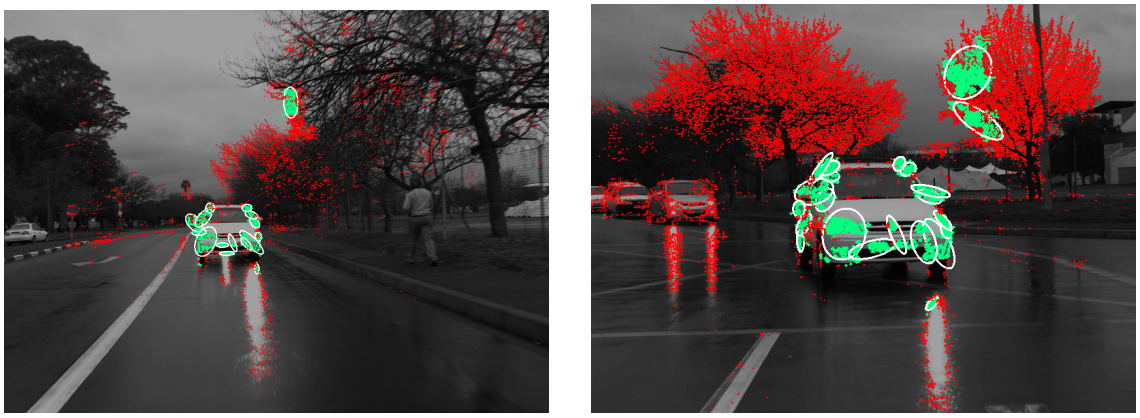


Figure 8.12: Evaluation results of the R44 sequence. (a) OSPA error of the GIW-PHD filter hypotheses for the track fusion algorithm and for vision-only processing. The sudden peak near the 80$^{\text{th}}$ time step is due to the missed detection of a pedestrian. (b) The corresponding cardinality results. (c) OSPA error of both the GIW-PHD filter hypotheses and the accompanying fused cluster measurement set. Note the improved localisation accuracy induced by filtering. (d) OSPA error of the cluster measurement set passed to the GIW-PHD filter for both the fusion and no fusion scenarios.

similar to the preceding datasets, however many clutter measurements originate from the trees on the side of the road. A vehicle and pedestrian also appear near the end of the sequence.



(a) (b)

Figure 8.13: Missed detection in the Merriman sequence. (a) High frame-to-frame distance impedes the performance of the feature tracker. (b) The second nearest vehicle is included in the ground truth, but is too far from the sensing platform to be reliably detected. Note also the prevalence of clutter detections that originate from foliage.

It is apparent that the OSPA error values shown in Figure 8.14 are considerably higher than those of the previous sequences. The poor performance can be explained with the help of the cardinality plot shown in Figure 8.14b, from which it can be seen that neither the fusion pipeline nor camera-only tracking is able to correctly estimate the number of targets. In the first half of the sequence, numerous unwanted stationary clutter objects are detected. These measurements are passed to the GIW-PHD tracker, leading to an increased cardinality estimate. The tracking of clutter measurements is the main reason for the large OSPA error in the sequence.

Two objects also proceed completely undetected, as shown in Figures 8.13a and 8.13b. For the first target, high frame-to-frame movement impedes the performance of the feature tracker. The second is a fair distance from the sensor platform, where the vision system often tends to fail. The radar also failed to detect any target in these particular frames. Over the duration of the Merriman sequence, only a single vehicle is reliably tracked.

The precision, recall, and MOTA values given in Table 8.3 suggest the complete failure of the GIW-PHD filter with regard to extent estimation. For both the fusion and non-fusion scenarios, these values are lower than what was achieved for the respective cluster measurement evaluations. However, the influence of clutter measurements explain these events. Recall that the two-dimensional track-to-track and cluster-to-cluster fusion procedure disregards height information. In the Merriman dataset, clutter that is situated above the vehicle is grouped with measurements that originate from the vehicle. An elongated extent estimate results from the wrong grouping. Once the clutter disappears, it takes numerous time steps for the estimate to adjust to a better fit, and it is during this time that the measurement hypotheses outperform the GIW-PHD filter hypotheses. Adjusting the intersection over union threshold to 0.3 leads to results that are consistent to the previous evaluations. The MOTA value remains low due to the high number of hypotheses that describe clutter.

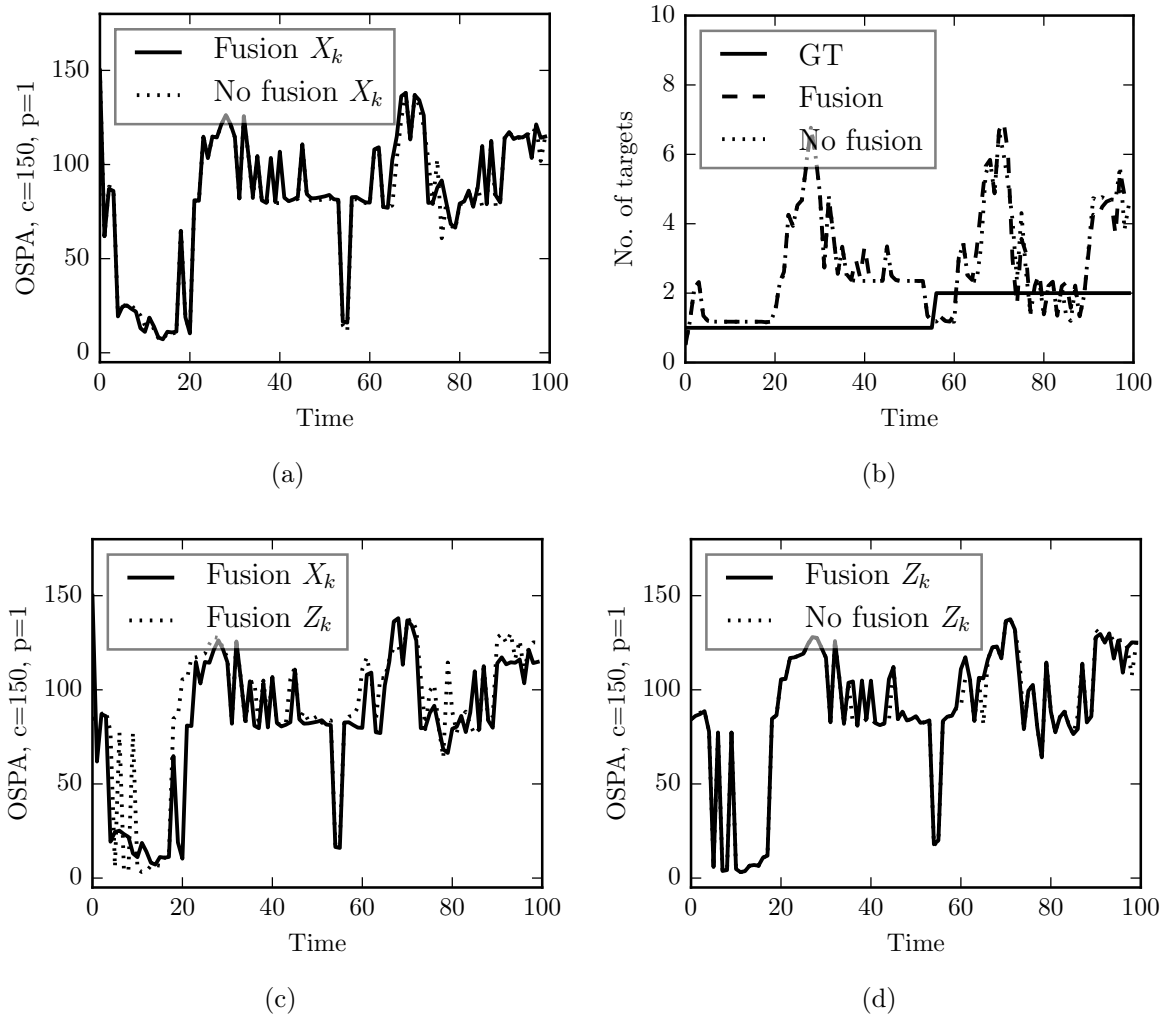


Figure 8.14: Evaluation results of the Merriman sequence. (a) OSPA error of the GIW-PHD filter hypotheses for the track fusion algorithm and for vision-only processing. (b) The corresponding cardinality results are adversely effected by clutter measurements, contributing to high OSPA values. (c) OSPA error of both the GIW-PHD filter hypotheses and the accompanying fused cluster measurement set. (d) OSPA error of the cluster measurement set passed to the GIW-PHD filter for both the fusion and no fusion scenarios.

Table 8.3: Tracking performance metrics for the Merriman sequence. Clutter interference impedes the system's performance.

	MOTP	MOTA	ID switches	Precision	Recall	$\overline{\text{OSPA}}$
GIW-PHD fusion	0.63	-1.37	0	0.09	0.15	82.3
GIW-PHD no fusion	0.61	-1.32	0	0.09	0.15	80.8
$Z_{\text{fusion},k}$ fusion	0.73	-0.61	N/A	0.34	0.65	88.8
$Z_{\text{fusion},k}$ no fusion	0.73	-0.56	N/A	0.35	0.65	87.7

8.5.5 Discussion

The practical results show that very accurate centre point tracking is achieved for persistent tracks. Pixel errors in the order of tens of pixels or less are consistently maintained when the tracker's cardinality estimate is correct. Moreover, the results show that the GIW-PHD filtering of the fused measurements lead to a steady increase in the localisation accuracy. The aforementioned testifies to the effectiveness of random matrix model, and proves the applicability of Bayesian extended target modelling to radar-vision fusion.

The primary factor contributing to periodic increases in the OSPA localisation error is the miss detection of targets and the formation of tracks on clutter. These events are mainly caused by the failure of either measurement extraction or data fusion. Sensor failures explain many of the missed detection events, with the inexpensive radar often failing completely. The effect of nearby clutter is also pronounced when radar state estimates are not available, since the second clustering routine allows less dense clusters.

The practical results do not provide conclusive evidence to suggest that track fusion outperforms vision-only processing. The sensing hardware prohibited the use of truly representative datasets, since reliable data could not be extracted in cluttered and/or multi-target environments. Datasets that enabled the testing of the algorithm most often included only a single target. Track fusion is expected to give more definitive performance gains in cluttered scenarios where the radar maintains its performance.

Chapter 9

Conclusion

This project was initiated with the goal of implementing an environmental perception system for detection and tracking of moving objects (DATMO). The ability to perceive the presence of moving objects in a robot's environment facilitates safe and reliable navigation, and is at the root of autonomous collision avoidance. The challenges brought forth in a DATMO environment lead to the decision to consider data fusion as a means to address the problem, with specific focus on the combination of radar and stereo vision sensors.

The traditional approach to DATMO was adopted, which follows the chronological steps of measurement extraction, data association and filtering. Algorithms that address measurement extraction was implemented to serve as a proof-of-concept for fusion and multi-target tracking. Measurements from the radar and vision subsystems were subsequently combined in a track-to-track data fusion algorithm, with the goal of achieving increased levels of perception performance. Extensive use was made of PHD filtering architectures in order to infer reliable information from the sensors' data, whilst circumventing the difficulties inherent to data association. Moreover, extended targets were accounted for by the adoption of appropriate dynamic and measurement models.

9.1 Summary

The following paragraphs summarise the methods that were implemented in the project, and include remarks on the results where applicable.

Chapter 1 introduces the DATMO problem and the need for data fusion to achieve robust perception performance. **Chapter 2** describes the concepts pertaining to the traditional DATMO processing chain. The chapter concludes with a review of the state-of-the-field with regard to radar-vision data fusion, and goes on to mention the lack of probabilistic extent modelling in previous work.

Chapter 3 presents the hardware configuration of the data fusion system and discusses some of the physical phenomena that govern measurement generation. It also describes the novel extrinsic sensor-to-sensor calibration method that was developed to determine the relative alignment of the respective sensors. The chosen hardware was successful in enabling the system to be demonstrated practically. However, the inexpensive radar often failed completely, which resulted in datasets unfit for the evaluation of data fusion in clutter or multi-target scenarios. The extrinsic calibration proved to be effective, yielding a reduced RMS error compared to a measurement procedure.

Chapter 4 describes the measurement extraction algorithms that were implemented for the vision and radar subsystems. The FMCW mode of operation was chosen for the radar, and the resulting data processed using standard two-dimensional Fourier analysis. A sparse motion-based technique coupled with density-based clustering was used to extract measurements from stereo vision image sequences. The algorithm was motivated by the requirement to extract information for as great a region of the image as possible, while also maintaining acceptable efficiency. Measurement extraction was not the project's primary focus, but it was required for practical evaluation purposes, for which the techniques proved sufficient.

Chapters 5 and 6 provide the theoretical background of the GM-PHD filter and the random matrix model respectively. Chapter 6 concludes with the presentation of the GIW-PHD filter, which is one of the main contributions of the work. The filter embeds the random matrix extended target model of Feldmann et al [41] in the GM-PHD filter of Vo and Ma [33]. The resulting GIW-PHD filter enables the use of sensor error models, contrary to an earlier GIW-PHD filter variant formulated by Granström and Orguner [76]. In Chapter 8, the proposed extended target PHD filter is evaluated against the point target GM-PHD filter by means of a two-dimensional simulation. It was found that the violation of the point target assumption is detrimental to the estimation performance of the GM-PHD filter, while the GIW-PHD filter maintained very low localisation and cardinality errors when served with the simulated cluster measurements.

Chapter 7 explains the proposed track-to-track data fusion architecture as well as the practical details that are required to realise the GM-PHD and GIW-PHD filters, namely mixture pruning and merging, and the formulation of the respective filters' mixture models. A mixture spawning strategy that adapts to regions where measurements occur is also presented.

Chapter 8 presents the evaluation results of various simulations as well as tests on real-world datasets. The extended target simulation environment is adapted in order to simulate the proposed track-to-track data fusion algorithm against the single-sensor case. The fusion simulation demonstrated that both the non-linear radar GM-PHD filter and the GIW-PHD filter provide reliable state estimates in high clutter MTT scenarios. In addition, the simulation also proved the effectiveness of the proposed formulation of target spawn mixtures. Simulation results indicated improved performance induced by track fusion as opposed to the two-stage clustering without radar state estimates. Fusion resulted in more accurate cluster measurements due to the method's increased resilience to clutter interference. Vision-only processing proved to be considerably more sensitive to clutter. The simulation also highlighted GIW-PHD filter's sensitivity to poorly clustered measurements: Compared to the earlier extended target simulation which received perfect clusters, a consistent error increase was witnessed using fusion-based clustering. However, the localisation error remained quite low despite the interference from clutter, missed detections, and measurement noise. Chapter 8 concludes with the analysis of various entities of the algorithm on the practical data. The results showed that the GIW-PHD filtering of the fused measurements lead to a consistent increase in the localisation accuracy, thereby reinforcing the value of state estimation. Quantitative analysis indicated very accurate centre point tracking using the extended target model, with errors in the order of tens of pixels for persistent target tracks. Clutter measurements and missed detections were the main contributors to high OSPA errors. Matched GIW-PHD hypotheses achieved a fair intersection over union overlap score, proving the sufficiency of random matrix extent

modelling. The practical results did not provide conclusive evidence to suggest that track fusion outperforms vision-only processing. The sensing hardware prohibited testing in cluttered environments, since reliable radar estimates could not be extracted from such data. Fusion is still expected to lead to improved performance in cluttered scenarios where the radar maintains its performance.

9.2 Contributions

The following list contains the most important contributions that originated from the research.

1. The most important contribution of this research is the PHD filter implementation of Feldmann et al.'s [41] random matrix extended target model. The filter explicitly accounts for extended targets, while allowing for a Gaussian representation of uncertainty for both target kinematic states and sensor error. Furthermore, multiple extended targets can be tracked simultaneously without explicit data association. A previous GIW-PHD filter has been implemented by Granström and Orguner [76], but the formulation does not allow the use of sensor error models. The proposed filter is expected to lead to increased performance in scenarios where the sensor noise is comparable to object extent.

2. The mixture spawning formulation used in both the GM-PHD and GIW-PHD filters represent a convenient strategy to spawn new target mixtures. The corresponding examples given by Vo and Ma [33] in their presentation of the GM-PHD filter rely on stationary mixtures that coincide with areas where targets are expected to originate from. The proposed spawn mixture's use of measurements enable it to generalise to many tracking scenario. Moreover, it requires considerably less assumptions with regard to the environment. A similar approach is presented by Yang et al. [84], but the method is designed for a particle PHD filter.

3. The explicit consideration of a Bayesian extended target model for fused radar and vision measurements promotes the credibility of the eventual state estimates. The implementation is not novel in itself, but related research demonstrates a neglect of appropriate extent modelling.

4. The extrinsic radar-to-stereo vision sensor calibration serves as the final contribution. The technique provided satisfactory results of parameters that would have been difficult to measure accurately, e.g. the antenna baseline. The setup is also simple and easy to implement.

9.3 Future Work

Potential research opportunities that could follow from this work will now be discussed.

1. The primary recommendation relates to the random matrix measurement model of the GIW-PHD filter. The update expects perfectly clustered measurements, which is difficult to achieve in cluttered environments. Imperfect clustering was seen to be one of the dominant causes of failure for the GIW-PHD filter. The author recommends analysis into a 'disjoint' measurement model that is formulated to

infer kinematic and extent information from numerous cluster measurements that are spread over the surface of the target. Over-segmented clusters can then be passed to the filter without prior grouping. It is much easier to obtain an over-segmentation than a proper segmentation. The realisation of such a filter would hold great promise for extended target tracking in cluttered environments. The first step toward the implementation of an over-segmented extended target random matrix measurement model would relate to the adaptation of the GIW mixture reduction equations. Mixture-to-mixture merging thresholds should then be chosen so as to force mixtures scatted over an object to merge into a single mixture.

2. Probabilistic fusion in the GIW-PHD filter holds potential for robust environmental perception. Using sensors that are more alike, such as an electronically scanned radar and 3D lidar, may work well when combined in the random matrix framework, since the measurements from both subsystems would resemble extended target measurements. It would, however, be necessary to derive non-linear random matrix recursive update equations to track such measurements in Cartesian coordinates. To the best of the author's knowledge, no such work exists.

3. A dense scene flow algorithm that combines stereo vision and radar information may prove fruitful. Radar measurements may be able to assist scene flow estimation for low textured image regions. These regions are difficult to match across different frames in vision-only scene flow algorithms.

4. The final recommendation relates to mixture spawning. If more sensitive radar hardware is used, the performance may be increased by making the mixture spawning more reliant on radar velocity information. Fewer tracks should then form from vision-based clutter .

Appendices

Appendix A

Data Fusion Algorithm

This chapter contains the track-to-track data fusion algorithm of Section 7.3. The algorithm was described in the relevant section, but implementation details were not given. This chapter provides the algorithmic description of the track fusion algorithm, and should be read in conjunction with the description given in Section 7.3. The pseudo code for track-to-track data fusion algorithm is shown in Algorithm 2.

Algorithm 2 Track-to-track data fusion algorithm.

Require: The state set describing moving GM-PHD radar mixture components in the two-dimensional CRF $X_{\text{rdr},k}^{C,\text{2-D}} = \{w_k^{(i)}, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)}\}_{i=1}^{J_{\text{rdr},k}}$, the set of two-dimensional CRF vision cluster measurements $Z_{\text{cam},k}^{C,\text{2-D}} = \{Z_k^{(i)}\}_{i=1}^{N_{\text{cam},k}}$, and a minimum distance threshold T.

▷ Associate camera cluster measurements to radar mixture components
$M \leftarrow \left\{\emptyset^{(i)}\right\}_{i=1}^{J_{\text{rdr},k}}$
$C \leftarrow \{1, \cdots, N_{\text{cam},k}\}$
for all $i \leftarrow 1, N_{\text{cam},k}$ **do**
  $\mathbf{m}, \mathbf{P} \leftarrow \textsc{GaussianApproximate}(Z_k^{(i)})$
  $j \leftarrow \underset{j \in \{1,\cdots,J_{\text{rdr},k}\}}{\arg\min} \mathcal{B}(\mathbf{m}, \mathbf{P}, \mathbf{m}_k^{(j)}, \mathbf{P}_k^{(j)})$
  $d \leftarrow \mathcal{B}(\mathbf{m}, \mathbf{P}, \mathbf{m}_k^{(j)}, \mathbf{P}_k^{(j)})$
  if $d < T$ **then**
    $M^{(j)} \leftarrow M^{(j)} \cup i$
    $C \leftarrow C \backslash i$
  end if
end for

▷ Build fused measurement set
▷ 1: Add track-fused measurements
$Z_{\text{fusion},k} = \emptyset$
for all $i \leftarrow 1, J_{\text{rdr},k}$ **do**
  $Z \leftarrow \{\mathbf{z}^{(n)}\}_{n=1}^{N} \leftarrow \textsc{SampleMixture}(w_k^{(i)}, \mathbf{m}_k^{(i)}, \mathbf{P}_k^{(i)})$
  for all $j \in M^{(i)}$ **do**
    $Z \leftarrow Z \cup Z_k^{(j)}$
  end for
  $Z_{\text{fusion},k} \leftarrow Z_{\text{fusion},k} \cup \{Z\}$
end for
▷ 2: Recluster non-grouped vision clusters and add to $Z_{\text{fusion},k}$
$Z \leftarrow \emptyset$
for all $i \in C$ **do**
  $Z \leftarrow Z \cup Z_k^{(i)}$
end for
$Z' \leftarrow \{Z'^{(n)}\}_{n=1}^{N} \leftarrow \text{DBSCAN}(Z)$
$Z_{\text{fusion},k} \leftarrow Z_{\text{fusion},k} \cup Z'$
$Z_{\text{fusion},k} \leftarrow \textsc{CameraToWorld}(Z_{\text{fusion},k})$ ▷ Convert back to WRF
return $Z_{\text{fusion},k}$

Appendix B

Rejection Sampling

Rejection sampling, or accept-reject sampling, uses a proposal probability distribution that serves to simulate a given target distribution [85]. The simulations described in Chapter 8 requires uniformly sampled ellipses. To sample from uniformly from the surface of a two-dimensional ellipse, a uniform two-dimensional rectangular distributions is used as the proposal distribution. Points are sampled one by one, and any that falls outside the surface of the ellipse is rejected. The procedure concluded when the required number of points have been accepted. The pseudo code for rejection sampling is shown in Algorithm 3. The result is a non-rotated ellipse, centred on the origin. The points can subsequently be rotated and transformed to represent an arbitrary two-dimensional ellipse.

Algorithm 3 Rejection sampling algorithm for sampling points uniformly from the surface of an ellipse.

Require: The ellipse parameters in the form of the major axis a, minor axis b, and the required number of samples N.

$P \leftarrow \emptyset$

while $|P| < N$ **do**

▷ Sample from uniform distribution $\mathcal{U}$

  $x \leftarrow 2a \sim \mathcal{U}(-0.5, 0.5)$

  $z \leftarrow 2b \sim \mathcal{U}(-0.5, 0.5)$

  if $(x/a)^2 + (z/b)^2 < 1$ **then**

    $P \leftarrow P \cup [x, z]^T$

  end if

end while

return P

Bibliography

[1] A. Azim and O. Aycard, "Detection, classification and tracking of moving objects in a 3d environment," in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. IEEE, 2012, pp. 802–807.

[2] D. F. Wolf and G. S. Sukhatme, "Mobile robot simultaneous localization and mapping in dynamic environments," *Autonomous Robots*, vol. 19, no. 1, pp. 53–65, 2005.

[3] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *Int. J. Rob. Res.*, vol. 26, no. 9, pp. 889–916, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1177/0278364907081229

[4] A. Petrovskaya, M. Perrollaz, L. Oliveira, L. Spinello, R. Triebel, A. Makris, J.-D. Yoder, C. Laugier, U. Nunes, and P. Bessiere, "Awareness of road scene participants for autonomous driving," in *Handbook of Intelligent Vehicles*. Springer, 2012, pp. 1383–1432.

[5] M. Mallick, V. Krishnamurthy, and B.-N. Vo, *Integrated tracking, classification, and sensor management: theory and applications*. John Wiley & Sons, 2012.

[6] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, Jul 1983.

[7] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, Dec 1979.

[8] A. Petrovskaya and S. Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Autonomous Robots*, vol. 26, no. 2, pp. 123–139, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10514-009-9115-1

[9] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[10] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, Jan 1997.

[11] H. Durrant-Whyte and T. C. Henderson, "Multisensor data fusion," in *Springer Handbook of Robotics*. Springer, 2008, pp. 585–610.

[12] M. Richards, J. Scheer, J. Scheer, and W. Holm, *Principles of modern radar*, ser. Principles of Modern Radar. SciTech Publishing, Incorporated, 2010, no. v. 1.

[13] B. Clarke, S. Worrall, G. Brooker, and E. Nebot, "Towards mapping of dynamic environments with fmcw radar," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013, pp. 147–152.

[14] M. Brown, "Feature extraction techniques for recognizing solid objects with an ultrasonic range sensor," *IEEE Journal on Robotics and Automation*, vol. 1, no. 4, pp. 191–205, 1985.

[15] J. J. Leonard and H. F. Durrant-Whyte, *Directed Sonar Sensing for Mobile Robot Navigation*. Norwell, MA, USA: Kluwer Academic Publishers, 1992.

[16] R. H. Rasshofer, M. Spies, and H. Spies, "Influences of weather phenomena on automotive laser radar systems," *Advances in Radio Science*, vol. 9, pp. 49–60, Jul. 2011.

[17] N. Paragios and G. Tziritas, "Adaptive detection and localization of moving objects in image sequences," *Signal Processing: Image Communication*, vol. 14, no. 4, pp. 277–296, 1999.

[18] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1219–1225.

[19] E. D. Dickmanns, *Dynamic Vision for Perception and Control of Motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[20] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of Mathematical Models in Computer Vision*. Springer, 2006, pp. 237–257.

[21] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/kanade meets horn/schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000045324.43199.43

[22] J. yves Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[23] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11263-010-0390-2

[24] J. Klappstein, T. Vaudrey, C. Rabe, A. Wedel, and R. Klette, *Moving object segmentation using optical flow and depth information*. Springer, 2009.

[25] A. Wedel, A. Meißner, C. Rabe, U. Franke, and D. Cremers, "Detection and segmentation of independently moving objects from dense scene flow," in *Energy minimization methods in computer vision and pattern recognition*. Springer, 2009, pp. 14–27.

[26] A. Giachetti, M. Campani, and V. Torre, "The use of optical flow for road navigation," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 1, pp. 34–48, 1998.

[27] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion*. Norwood, MA, USA: Artech House, Inc., 2007.

[28] H. de Waard, "A new approach to distributed data fusion," Ph.D. dissertation, The University of Amsterdam, 2008.

[29] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Multi-target tracking using joint probabilistic data association," in *Decision and Control including the Symposium on Adaptive Processes, 1980 19th IEEE Conference on*, Dec 1980, pp. 807–812.

[30] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2008.

[31] H. Durrant-Whyte, "Multi sensor data fusion." Australian Centre for Field Robotics, 2001.

[32] R. P. S. Mahler, "Multitarget bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, Oct 2003.

[33] B. N. Vo and W. K. Ma, "The gaussian mixture probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, Nov 2006.

[34] K. Granström and M. Baum, "Extended object tracking: Introduction, overview and applications," *CoRR*, vol. abs/1604.00970, 2016.

[35] L. Mihaylova, A. Y. Carmi, F. Septier, A. Gning, S. K. Pang, and S. Godsill, "Overview of bayesian sequential monte carlo methods for group and extended object tracking," *Digital Signal Processing*, vol. 25, pp. 1 – 16, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200413002716

[36] K. Gilholm and D. Salmond, "Spatial distribution model for tracking extended objects," *IEE Proceedings - Radar, Sonar and Navigation*, vol. 152, no. 5, pp. 364–371, October 2005.

[37] . Baum and . D. Hanebeck, "Random hypersurface models for extended object tracking," in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec 2009, pp. 178–183.

[38] M. Baum and U. D. Hanebeck, "Shape tracking of extended objects and group targets with star-convex rhms," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, July 2011, pp. 1–8.

[39] J. W. Koch, "Bayesian approach to extended object and cluster tracking using random matrices," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, no. 3, pp. 1042–1059, July 2008.

[40] M. Baum, M. Feldmann, U. D. Hanebeck, and W. Koch, "Extended object and group tracking: A comparison of random matrices and random hypersurface models," in *In Proceedings of the IEEE ISIF Workshop on Sensor Data Fusion*, 2010.

[41] M. Feldmann, D. Franken, and W. Koch, "Tracking of extended objects and group targets using random matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1409–1420, April 2011.

[42] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, Dec. 2006. [Online]. Available: http://doi.acm.org/10.1145/1177352.1177355

[43] T.-D. Vu, O. Aycard, and N. Appenrodt, "Online localization and mapping with moving object tracking in dynamic outdoor environments," in *2007 IEEE Intelligent Vehicles Symposium*, 2007.

[44] Y. Bar-Shalom and X.-R. Li, "Multitarget-multisensor tracking: principles and techniques," *Storrs, CT: University of Connecticut, 1995.*, 1995.

[45] A. Gern, U. Franke, and P. Levi, "Robust vehicle tracking fusing radar and vision," in *Multisensor Fusion and Integration for Intelligent Systems, 2001. MFI 2001. International Conference on*, 2001, pp. 323–328.

[46] T. Wang, N. Zheng, J. Xin, and Z. Ma, "Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications," *Sensors*, vol. 11, no. 9, pp. 8992–9008, 2011.

[47] Z. Ji and D. Prokhorov, "Radar-vision fusion for object classification," in *Information Fusion, 2008 11th International Conference on*, June 2008, pp. 1–7.

[48] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 95–105, March 2007.

[49] S. Wu, S. Decker, P. Chang, T. Camus, and J. Eledath, "Collision sensing by stereo vision and radar sensor fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 4, pp. 606–614, Dec 2009.

[50] Y. Fang, I. Masaki, and B. Horn, "Depth-based target segmentation for intelligent vehicles: fusion of radar and binocular stereo," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 3, pp. 196–202, Sep 2002.

[51] E. Richter, R. Schubert, and G. Wanielik, "Radar and vision based data fusion - advanced filtering techniques for a multi object vehicle tracking system," in *Intelligent Vehicles Symposium, 2008 IEEE*, June 2008, pp. 120–125.

[52] F. Garcia, P. Cerri, A. Broggi, A. De la Escalera, and J. M. Armingol, "Data fusion for overtaking vehicle detection based on radar and optical flow," in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. IEEE, 2012, pp. 494–499.

[53] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[54] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[55] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

[56] J.-Y. Bouguet, "Camera calibration toolbox for matlab," 2004.

[57] M. I. Skolnik, *Introduction to Radar Systems*. McGraw Hill, 2001.

[58] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, Jun 1994, pp. 593–600.

[59] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944. [Online]. Available: http://www.jstor.org/stable/43633451

[60] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[61] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 430–443.

[62] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls*, 1997.

[63] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. part i. dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, Oct 2003.

[64] J. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.

[65] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." AAAI Press, 1996, pp. 226–231.

[66] D. E. Barrick, "Fm/cw radar signals and digital processing," DTIC Document, Tech. Rep., 1973.

[67] B. J. Lipa and D. E. Barrick, "Fmcw signal processing," *FMCW signal processing report for Mirage Systems*, 1980.

[68] G. V. Trunk, "Range resolution of targets using automatic detectors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-14, no. 5, pp. 750–755, Sept 1978.

[69] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range fmcw monopulse radar for hand-gesture sensing," in *2015 IEEE Radar Conference (RadarCon)*, May 2015, pp. 1491–1496.

[70] B. T. Vo, "Random finite sets in multi-object filtering," Ph.D. dissertation, University of Western Australia, 2008.

[71] K. Granström and U. Orguner, "On the reduction of gaussian inverse wishart mixtures," in *Information Fusion (FUSION), 2012 15th International Conference on.* IEEE, 2012, pp. 2162–2169.

[72] S. W. Nydick, "The wishart and inverse wishart distributions," May 2012, last visited on 21/10/2016. [Online]. Available: http://www.tc.umn.edu/~nydic001/docs/unpubs/Wishart_Distribution.pdf

[73] S. M. Lynch, *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media, 2007.

[74] R. Mahler, "Phd filters for nonstandard targets, i: Extended targets," in *Information Fusion, 2009. FUSION '09. 12th International Conference on*, July 2009, pp. 915–921.

[75] E. W. Weisstein, "Set partition. From MathWorld—A Wolfram Web Resource," last visited on 21/10/2016. [Online]. Available: http://mathworld.wolfram.com/SetPartition.html

[76] K. Granstrom and U. Orguner, "A phd filter for tracking multiple extended targets using random matrices," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5657–5671, Nov 2012.

[77] K. T. Abou-Moustafa and F. P. Ferrie, "A note on metric properties for some divergence measures: The gaussian case." in *ACML*, 2012, pp. 1–15.

[78] "Technical Application Note (TAN2004006): Stereo Accuracy and Error Modeling," Point Grey, Tech. Rep., Aug 2012.

[79] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/246309

[80] D. Schuhmacher, B. T. Vo, and B. N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, Aug 2008.

[81] K. Granstrom, C. Lundquist, and O. Orguner, "Extended target tracking using a gaussian-mixture phd filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3268–3286, October 2012.

[82] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vision*, vol. 101, no. 1, pp. 184–204, Jan. 2013.

[83] L. Leal-Taixé, A. Milan, I. D. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *CoRR*, vol. abs/1504.01942, 2015.

[84] F. Yang, Y. Wang, H. Chen, P. Zhang, and Y. Liang, "Adaptive collaborative gaussian mixture probability hypothesis density filter for multi-target tracking," *Sensors*, vol. 16, no. 10, p. 1666, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/10/1666

[85] G. Casella, C. P. Robert, and M. T. Wells, "Generalized accept-reject sampling schemes," *Lecture Notes-Monograph Series*, pp. 342–347, 2004.